ELSEVIER

Cognitive Science 27 (2003) 561–590

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

A real-world rational agent: unifying old and new AI

Paul F.M.J. Verschure [a,*], Philipp Althaus [b]

[a] *Institute of Neuroinformatics, University/ETH Zürich, Zürich, Switzerland*
[b] *Centre for Autonomous Systems, Royal Institute of Technology, S-10044 Stockholm, Sweden*

Received 9 January 2002; received in revised form 8 August 2002; accepted 10 December 2002

Abstract

Explanations of cognitive processes provided by traditional artificial intelligence were based on the notion of the knowledge level. This perspective has been challenged by new AI that proposes an approach based on embodied systems that interact with the real-world. We demonstrate that these two views can be unified. Our argument is based on the assumption that knowledge level explanations can be defined in the context of Bayesian theory while the goals of new AI are captured by using a well established robot based model of learning and problem solving, called Distributed Adaptive Control (DAC). In our analysis we consider random foraging and we prove that minor modifications of the DAC architecture renders a model that is equivalent to a Bayesian analysis of this task. Subsequently, we compare this enhanced, "rational," model to its "non-rational" predecessor and a further control condition using both simulated and real robots, in a variety of environments. Our results show that the changes made to the DAC architecture, in order to unify the perspectives of old and new AI, also lead to a significant improvement in random foraging.

Keywords: Artificial intelligence; Neuroscience; Psychology; Cognitive architecture; Decision making; Intelligent agents; Learning; Machine learning; Problem solving; Situated cognition; Computer simulation; Neural networks; Robotics

1. Introduction

Traditional AI aims to explain intelligent behavior at the knowledge level (Newell, 1982). The knowledge level describes intelligent behavior in terms of knowledge, goals and actions.

* Corresponding author. Tel.: +41-1-635-3031; fax: +41-1-635-3053.
E-mail address: pfmjv@ini.phys.ethz.ch (P.F.M.J. Verschure).

0364-0213/03/$ – see front matter © 2003 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/S0364-0213(03)00034-X

Knowledge and goals are organized following the principle of rationality: "... if the system wants to attain goal *G* and knows that to do act *A* will lead to attaining *G*, then it will do *A*. This law is a simple form of rationality that an agent will operate in its own best interest according to what it knows" (Newell, 1990). The empirical hypothesis put forward in this approach is that general intelligence can only be displayed by systems that can manipulate symbols: i.e. physical symbol systems (Newell, 1980; Newell & Simon, 1976). This view has been criticized on several grounds and a number of fundamental problems have been identified; the frame problem (McCarthy & Hayes, 1969), the symbol grounding problem (Harnad, 1990; Searle, 1982), the frame of reference problem (Clancey, 1989a), and the problem of situatedness (Suchman, 1987) (see Pfeifer & Scheier, 1999 for a review). It has been argued that most of these problems can be brought back to the critical dependence of the proposed solutions on the *a priori* specification of rules and representations, the problem of priors (Verschure, 1998). Against this background, so-called new AI has emerged, emphasizing the importance of situatedness and grounding through the use of real-world systems, i.e. robots (Brooks, 1991a). Proponents of this view have argued that explanations of intelligent behavior can be found without relying on internal symbolic representations (Brooks, 1991b) or of goals (Pfeifer, 1995). This change of perspective in AI raises the important question whether these two views on intelligence are incompatible or whether they can be unified (Verschure, 1993). One motivation for trying to unify these two views is that they both seem to capture different aspects of intelligence. Where the traditional approach found effective descriptions of higher-level cognitive processes, such as problem solving and planning (Newell, 1990), new AI has aimed to solve problems in the real-world, incorporating more biologically motivated principles in its solutions (Pfeifer & Scheier, 1999). However, traditional AI failed to ground its solutions in the real-world, while new AI faces the challenge to scale up to non-trivial cognitive processes.

In this paper we aim at bridging the apparent gap between the perspectives of old and new AI. We show that these two views on intelligence can be unified, provided one is willing to accept specific definitions of the knowledge level and of a situated agent and its control structure. Our proposal is based on the assumption that a knowledge level description of intelligence, including the principle of rationality, can be captured in the perspective of Bayesian decision making (Bayes, 1763). We satisfy, in parallel, the goals of new AI by using a well established robot based model of learning and problem solving, called Distributed Adaptive Control (DAC) (Verschure, Kröse, & Pfeifer, 1992; Verschure & Voegtlin, 1998). In this paper we prove that DAC is equivalent to an optimal decision making system in a Bayesian sense. Most importantly we show that our solution is self-contained in the sense that DAC acquires and updates its own set of prior hypotheses. This is relevant since, as traditional AI, also a Bayesian framework does not automatically solve the symbol grounding problem: it also assumes that the knowledge of a decision making system is defined *a priori*. In order to prove that the DAC architecture obeys the principle of rationality, it needed to be modified. Using experiments with simulated and real robots we demonstrate that this modified model, called DAC5, shows better performance in a random foraging task than both its predecessor, DAC3, and a further control condition.

This paper does not aim to introduce Bayesian decision theory to cognitive science. To the contrary, the last few years have seen a surge in interest in the use of these techniques (see Haddawy, 1999; Russell & Norvig, 1995 for an overview). Nor is it the point of this paper

to show that DAC is or is not a better approach towards behavior based robotics and new AI. Rather we want to show that the perspectives on intelligent systems offered by old and new AI can be unified. In our proposal these two views capture different, but complementary, aspects of intelligence. We argue that neither of these can be ignored in the search to understand this complex phenomenon.

1.1. A Bayesian interpretation of the knowledge level

The knowledge level is seen as one of the levels at which intelligent systems must be described. The knowledge level, in turn, is implemented by the symbol, or program, level which is implemented at the hardware level. Where the knowledge level describes the competence of an intelligent system its architecture is the instantiated physical symbol system (Newell, 1980; Newell & Simon, 1976). This physical symbol system should approximate the competencies defined by its knowledge level specification. Hence, the knowledge level specifies the functional properties of the actually physically instantiated intelligent system in terms of knowledge, goals, and actions. On one hand the knowledge level can be seen as expressing an observer stance towards an intelligent system where knowledge and goals are attributed to an artificial system (Chandrasekaran, 1994; Clancey, 1996; Newell, 1982). On the other hand, the distinction between knowledge, goals, and actions does define key properties of the physical symbol system that has to satisfy a knowledge level specification.

The prototypical example of a physical symbol system that aims at satisfying knowledge level constraints is the SOAR-architecture proposed by Newell (Newell, 1990, 1992) (see Laird & Rosenbloom, 1996; Vinkhuyzen & Verschure, 1994; for a review). SOAR comes out of a tradition of modeling which spans over 30 years and started with the logical theorist proposed in the fifties followed by the general problem solver (Newell, Shaw, & Simon, 1959; Newell & Simon, 1963, 1972). All knowledge in SOAR is represented by productions (if–then constructs) and its working memory contains the current state of the problem solving process and its context. All productions freely add sentences to working memory until no more productions apply (SOAR runs into quiescence). During this elaboration phase preferences for the operators are accumulated in working memory. Preferences express whether a particular operator should be applied or not. After SOAR has run to quiescence, a decision procedure selects the operator that comes out most favorably in relation to the current goal and applies it to the current problem state to generate the next state. Hence, the SOAR-architecture selects operator after operator until it arrives at the goal state. However, it is quite possible that SOAR cannot decide what operator to use. In these cases, SOAR reaches an impasse and a new problem space, whose goal is the resolution of the impasse, is created. Within this sub-space, SOAR continues with the same decision cycle as before, it applies its operators until the goal of the sub-space is reached. This in turn leads to the selection of an operator in the problem space where the impasse originally occurred. The resolution of an impasse invariably leads to the creation of a chunk. A chunk is a regular production, which is added to the production memory. The condition-side of a chunk consists of the states that were true before the impasse occurred, and on the action-side of the chunk the operators that lead to the resolution of the impasse. This chunk will become active, whenever the situation that caused the impasse occurs again. The formation of chunks is seen as a form of learning that prevents SOAR from having to solve the same problem twice.

The knowledge level attributes knowledge, goals, and actions to an intelligent system which are in turn explicitly represented in a physical symbol system (Newell, 1990). The principle of rationality specifies that those actions are chosen that allow the system to achieve its goals. The Bayesian framework can be seen as an alternative description of the knowledge level utilizing the same concepts. In the Bayesian case knowledge is defined by a set of prior hypotheses S and the theorem of inverse probability defines the probability that hypothesis s is true given observation r:

$$p(s|r) = \frac{p(r|s)\,p(s)}{p(r)} \tag{1}$$

where $p(r)$ is the probability of making observation r, $p(s)$ the prior probability of s being true, and $p(r|s)$ the prior probability that making observation r given s is true. The optimal action, a, can be calculated using a score function $G_g(s_n,a)$ that defines the expected gain, $\langle g\rangle_a$, of performing action a given hypothesis s_n:

$$\langle g\rangle_a = \sum_{s_n \in S} p(s_n|r) G_g(s_n, a) \tag{2}$$

Bayes' principle states that optimal decision making requires that the action a_* is selected that maximizes the expectancy $\langle g\rangle$:

$$\langle g\rangle_{a_*} = \sum_{s_n \in S} p(s_n|r) G_g(s_n, a_*) = \max_{a_k \in A} \left[\sum_{s_n \in S} p(s_n|r) G_g(s_n, a_k) \right] \tag{3}$$

The rational view of decision making, also expressed in a Bayesian framework, has been widely applied in traditional AI to describe high-level cognitive functions (Newell, 1990). Since we want to unify this rational view with one of a situated real-world system using a Bayesian framework, an important question is whether animal and human behavior can be accurately described in this perspective. Many behavioral experiments have been performed to assess the optimality of animal decision making (Gallistel, 1990). A typical example are foraging experiments using mazes. For instance, rats placed in a radial arm maze, where different arms contain a varying amount of food pellets, develop an optimal foraging strategy in terms of travel time, probability of food occurrence and amount of food that adapts to changes in these task parameters (Roberts, 1992). It has been shown that the strategies adopted are strongly controlled by the expected gain and its magnitude (Herrnstein, 1970) and maintain an optimal balance between exploration and exploitation (Krebs, Kacelnik, & Taylor, 1978). Other examples can be found in human psychophysics, in particular using visual tasks (Knill & Richards, 1996; Weiss & Adelson, 1998). For instance, the perception of three-dimensional objects (Nakayama and Shimojo, 1992), of stereo vision (Porrill, Frisby, Adams, & Buckley, 1999), the integration of multiple, possibly ambiguous, measurements of local image properties in the perception of motion (Weiss & Adelson, 1998) and the use of prior knowledge in the detection of visual stimuli in static noise (Burgess, 1985). An important problem in perception is how multiple sources of information are integrated in complex recognition tasks, e.g. auditory and visual cues in speech perception. A comparison of different theoretical approaches towards this problem showed that a Bayesian model provided the most accurate description

and prediction of human performance (Massaro & Friedman, 1990). On the basis of these observations it has been proposed that Bayesian inference is a general organizing principle of perception and multi-model integration (Massaro, 1997). The above experiments characterized the performance of humans and animals in a Bayesian framework. But also several proposals have been made which attempt to describe the information processing performed by the neuronal substrate in this decision theoretic framework. For instance, it has been shown that neurons in the parietal cortex of monkeys accurately represent key decision making variables such as expected gain and gain magnitude (Platt & Glimcher, 1999). In other work it has been argued that a Bayesian framework can accurately describe the properties of the visual cortex in motion processing (Koechlin, Anton, & Burnod, 1999) and the multi-sensor fusion performed by the superior colliculus that allow it to trigger saccadic eye movements (Anastasia, Patton, & Belkacem-Boussaid, 2000).

The above examples do not prove that humans and animals are optimal Bayesian machines. In many tasks optimal decision making is biased by different psychological factors which might fall outside such a framework (Tversky & Kahneman, 1981; see Mellers, Schwartz, & Cooke, 1998 for a review). For example, it has been shown that the illusion to control the outcome of a choice task leads to an increase in the subjective estimate of the probability of success as compared to its objective probability (Fong & McCabe, 1999). In another set of experiments it was shown that humans use a variety of decision rules and that human subjects are not optimal in a Bayesian sense (El-Gamal & Grether, 1995). However, the same study showed that Bayes' rule is the one decision rule which approximates human performance closest. Although these experiments question the generality of the Bayesian framework, our earlier examples show that this framework does have experimental support in both psychology and neuroscience. Moreover, one could argue that the observed deviations from optimal decision making can be accounted for by adjusting the score functions employed. Hence, our attempt to use Bayesian inference as an operational definition of the knowledge level and its principle of rationality places these concepts on a well elaborated empirical base.

The rationalistic view of decision making provided by a Bayesian framework describes intelligence from a purposive functional perspective (Chater & Oaksford, 1999). As a knowledge level description it does not provide any direct link to the more mechanistic and embodied perspective of new AI. In particular, the set of prior hypotheses, at the core of a Bayesian inference system, need to be acquired and maintained by the system itself and not provided by its designer in order for the ontology of the problem solving system to be grounded. The robot based model of learning and problem solving we investigate here, called DAC, does satisfy this requirement (Verschure, 1998; Voegtlin & Verschure, 1999). DAC has been developed as a model for the behavioral paradigms of classical and operant conditioning (Mackintosh, 1974). In particular, it addresses the question how a real-world system can acquire, retain, and express knowledge of its interaction with the world. DAC has been investigated using both simulated and real robots in foraging and block sorting tasks (Verschure & Voegtlin, 1998, 1999; Verschure et al., 1992; Verschure, Wray, Sporns, Tononi, & Edelman, 1995) and is a standard in the field of new artificial intelligence and behavior based robotics (Arkin, 1998; Clancey, 1996; Hendriks-Jansen, 1996; McFarland & Bosser, 1993; Pfeifer & Scheier, 1999). Hence, the unification of new and old AI pursued here requires that we demonstrate that the DAC architecture satisfies a Bayesian analysis of its task domain, i.e. random foraging.

The following sections introduce our extension of the DAC model, called DAC5, and subsequently prove that it is equivalent to a Bayesian inference system. We investigate the implications of the extensions made to the model and compare it to its predecessor, DAC3, and an additional control condition using both simulated and real robots in random foraging tasks.

2. The learning model—DAC5

DAC was developed as a robot based neuronal model of classical and operant conditioning (Mackintosh, 1974). In classical conditioning initially neutral stimuli, Conditioned Stimuli (CS), become able to trigger Conditioned Responses (CR) due to their correlation with motivational stimuli, Unconditioned Stimuli (US) (Pavlov, 1927). Presentation of a US causes an automatic response, Unconditioned Response (UR). After several simultaneous presentations of US and CS, the presentation of the CS alone will trigger a response similar to the UR, the CR. In a typical classical conditioning freezing experiment, for instance, a tone (CS) is presented paired with a footshock (US). Initially only the US induces a freezing response (UR), the cessation of ongoing behavior. After several paired presentations of CS and US, however, presentation of the CS alone will induce freezing (CR). In operant, or instrumental, conditioning the animal learns to associate its actions with particular states of the environment (Thorndike, 1911). In this case the US that results from an action, provides a reinforcement signal for the learning system. In a typical experiment in a so-called Skinner box an animal learns to press a lever (CR) to receive a food reward (US). DAC is constructed on the assumption that both phenomena reflect components which are closely coupled in the overall learning system (Verschure & Voegtlin, 1998). In DAC the assumption is made that in order to explain these forms of learning three strongly coupled layers of control need to be distinguished: reactive, adaptive, and contextual control (Fig. 1).

1. The *reactive control* layer provides the behaving system with a basic level of behavioral competence based on prewired reflexive relationships between simple sensory events (US) and actions (UR).
2. The *adaptive control* layer allows the system to develop representations of complex sensory events (CS) conditional on US events.
3. The *contextual layer* supports the formation of more complex representations of CS and CR events expressing their relationship in time.

2.1. Reactive and adaptive control

The reactive and adaptive control layers are based on the following assumptions (Fig. 1):

- USs of a particular type activate specific populations of neurons reflecting an internal state (IS), i.e. aversive ($US^- \rightarrow IS^-$) and appetitive ($US^+ \rightarrow IS^+$).
- Cells in IS will activate specific reflexive actions (UR), i.e. $IS^- \rightarrow$ avoidance and $IS^+ \rightarrow$ approach.
- Conflict resolution in action selection is resolved through a predefined interaction between the IS populations via an inhibitory unit (I).

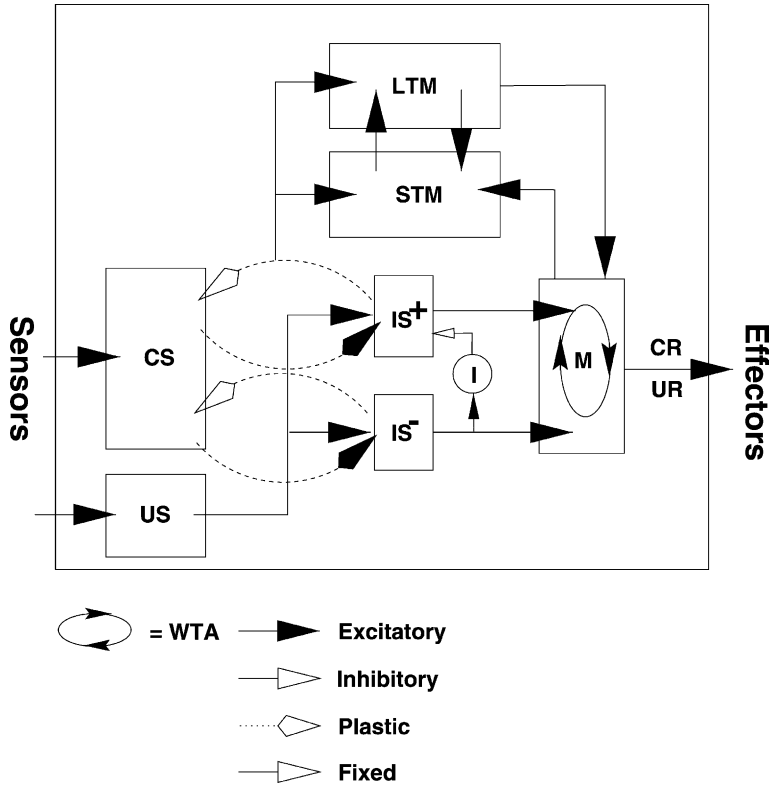

Fig. 1. Schematic representation of the DAC architecture. See text for explanation.

- CSs are derived from events on distal sensors (e.g. color CCD camera), while USs are derived from proximal sensors (e.g. collision sensors).
- CS representations are formed by modifying the connections between the CS and IS populations.

2.1.1. Model equations of the fast dynamics of the reactive and adaptive control layers

The activity, u_j, of unit j in population CS is derived from the state, s_j, of element j of the related distal sensor, using a transduction function f.

$$u_j = f(s_j) \tag{4}$$

The activity of population CS is propagated to the IS populations through excitatory connections. The input, v_i^m, of cell i in IS population m is defined by:

$$v_i^m = \sum_{j=1}^{M^{\mathrm{CS}}} w_{ij}^m u_j + p_i^m - \gamma^m I \tag{5}$$

where M^{CS} is the size of the CS population, w_{ij}^m is the efficacy of the connection between CS cell j and IS cell i, p_i^m is the state of element i of US conveying sensor m, I is the activity of

the inhibitory unit I, and γ^m the gain of inhibition to IS population m. The activity, o_i^m, of cell i of IS population m is defined by:

$$o_i^m = H(v_i^m - \theta_i^m) \tag{6}$$

where θ_i^m is the activation threshold and H the Heaviside or step function. The input to the inhibitory unit I at time $t+1$, $v^{\mathrm{I}}(t+1)$, is derived from the activity of the aversive internal state population IS^-:

$$v^{\mathrm{I}}(t+1) = \alpha^{\mathrm{I}} v^{\mathrm{I}}(t) + \gamma^{\mathrm{I}} \sum_{j=1}^{M^{\mathrm{IS}^-}} o_j^{\mathrm{IS}^-}(t) \tag{7}$$

where α^{I} ($\alpha^{\mathrm{I}} < 1$) is the persistence and γ^{I} is the excitatory gain of I. The activity of I is determined using Eq. (6), thresholding with θ^{I}. The input, f_k, of unit k in the UR population is defined by:

$$f_k = \sum_{m=1}^{K} \sum_{i=1}^{M^m} y_{ki}^m o_i^m \tag{8}$$

where K denotes the number of IS populations, M^m the size of IS population m, and y_{ki}^m the strength of the connection between cell i of IS population m and cell k of the UR population. After updating their inputs the UR units compete in a Winner Take All (WTA) fashion. If the activity of the winning unit is suprathreshold it will induce a specific motor action. In case no motor unit is activated the control structure will trigger exploration behavior (forward translation). A system consisting only of the US–IS and the IS–UR mapping constitutes a reactive control structure.

2.1.2. Model equations of the slow dynamics of the adaptive control layer

After updating the input, v^m, of the IS populations (Eq. (5)), these populations in turn recurrently inhibit the CS population. The resultant activity, u'_j, of unit j in the CS population is now defined as:

$$u'_j = u_j - \gamma^r e_j \tag{9}$$

where γ^r is a gain factor modulating the effect of the recurrent inhibition and e_j is the recurrent prediction defined by:

$$e_j = \sum_{m=1}^{K} \sum_{i=1}^{M^m} \frac{w_{ij}^m v_i^m}{M^m} \tag{10}$$

e will be referred to as a *CS prototype*. The connections between unit j of the CS population and unit i of IS population m, w_{ij}^m, evolve according to:

$$\Delta w_{ij}^m = \eta^m v_i^m u'_j \tag{11}$$

where η^m defines the learning rate of the connections between the CS population and IS population m. The synaptic weights of the connections between the CS and IS populations are

prevented from attaining negative values. Given the effect of the recurrent inhibition (Eq. (9)) this learning method is referred to as *predictive Hebbian learning* (Verschure & Pfeifer, 1992; Verschure & Voegtlin, 1998). This learning rule is related to filtering theory and state estimation (Kalman, 1960), and similar approaches have been applied to modeling the cortical mechanisms of perceptual learning (Rao, 1999; Rao & Ballard, 1999). A biophysically realistic real-time model of this learning rule has been defined that accounts for the physiological changes observed in the primary auditory cortex during classical conditioning (Sanchez-Montanes, König, & Verschure, 2002; Sanchez-Montanes, Verschure, & König, 2000) in combination with a physiologically and anatomically constrained model of the cerebellum (Hofstoetter, Mintz, & Verschure, 2002).

The adaptive control layer will over time form a classification of its interaction with the environment in terms of CS events conditional to its internal states. These acquired CS prototypes on one hand allow the system to function as an adaptive controller and on the other form the representational building blocks for the construction of sequential representations by the contextual control layer.

2.2. Model equations of the contextual control layer

The contextual control layer of DAC5 (Fig. 1) is based on an earlier model, called DAC3 (Verschure, 2000; Verschure & Voegtlin, 1998) and is based on the following assumptions:

- Salient events are stored in short-term memory (STM).
- The content of STM is stored in long-term memory (LTM) when a goal state is reached.
- The content of LTM is matched against ongoing sensory events.
- Matching LTM elements, or segments, bias action selection of the motor population.
- Chaining through LTM sequences is achieved by biasing LTM matching.

DAC5 bootstraps itself from a stage of adaptive control to a stage of contextual control. This transition depends on the quality of the matching between predicted and actual CS events expressed by an internal *discrepency measure*, D. Matching is defined by the distance, $d(u, e)$, between the feedforward generated CS activity pattern, u (Eq. (4)), and the recurrent prediction, e (Eq. (10)):

$$d(u, e) = \frac{1}{M^{\mathrm{CS}}} \sum_{j=1}^{M^{\mathrm{CS}}} \left| \frac{u_j}{\max_{1 \le j' \le N} u_{j'}} - \frac{e_j}{\max_{1 \le j' \le N} e_{j'}} \right| \tag{12}$$

D evolves according to:

$$D(t+1) = \alpha^D D(t) + (1 - \alpha^D) d(u, e) \tag{13}$$

where α^D defines the integration time constant. D is a dynamic state variable which is internal to the learning system. It provides an estimate of the progression of learning at the level of adaptive control and will decrease if the constructed CS prototypes consistently match ongoing CS events. It will increase if expected CS events are violated. Once D falls below a *confidence threshold*, θ^D, DAC5 engages its general purpose learning system consisting of structures for STM and LTM. STM functions as a ring buffer, which stores pairs of CS representations and

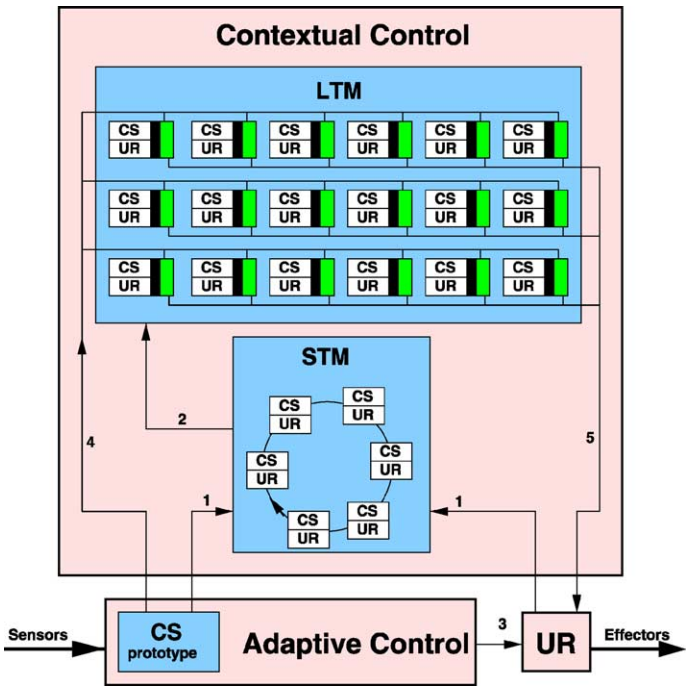

Fig. 2. The general purpose learning system of DAC5, which is constructed on top of the adaptive control layer. (1) The CS prototype and UR activity are written to the STM buffer and stored as a segment. (2) If a target or a collision occurs, the contents of STM are written to LTM as a sequence. Now each segment consists of the stored CS prototype, the stored UR activity, a trigger unit (black) and a collector unit (gray). (3) The UR population receives input from the IS populations according to the rules of the adaptive control structure. (4) If the input to UR is sub-threshold, the values of the current CS prototype are matched against those stored in LTM. (5) The collector units give an input to the UR population.

related motor actions (segments), and has a finite length, N^{STM}. LTM stores sequences of these segments (see Fig. 2 for illustration).

After each time step the generated CS prototype, e (Eq. (10)), and the action executed by the robot (Eq. (8) or (17) after WTA), are stored in the STM buffer as a segment. In case a goal state is reached, i.e. a target is found or a collision is suffered, the STM content is copied into LTM and STM is reset. This creates sequences "belonging" to different goal states: target found or collision suffered. If none of the IS populations is active (which means no action is triggered by the adaptive control layer), all the CS prototypes stored in the LTM segments will be matched against the current CS prototype (Eq. (10)). Matching of segment l of sequence q is expressed in the matching score m_{lq} defined by:

$$m_{lq} = d(e, g_{lq}) \tag{14}$$

where g_{lq} represents the CS prototype stored in segment l of LTM sequence q and d is defined in Eq. (12). The degree of matching of segment l in sequence q defines the input to its collector unit, c_{lq}:

$$c_{lq} = 1 - m_{lq} t_{lq} \tag{15}$$

where t_{lq}, is the activity of the trigger unit which will be introduced below. The activity of a collector unit, b_{lq} is defined as:

$$b_{lq} = H(c_{lq} - \theta^C)c_{lq} \tag{16}$$

where θ^C is the activation threshold. All collector units are connected to the UR population. The input to cell k in the UR population, f_k, receives input from all the collector units of segments that have stored the same action as represented by this unit:

$$f_k = \sum_{l \in \mathrm{LTM}} \pm \frac{b_{lq}}{z_{lq}} \tag{17}$$

where z_{lq} is the distance, measured in segments, between segment l to the end of its sequence or, in other words, the distance to the goal state. By dividing the output of segment lq with z_{lq} the segments closer to the goal state are weighted higher. The sign is plus if segment lq is from a sequence that was stored when a target was found and minus if it was stored when a collision was suffered. After updating their input, the UR units compete in a WTA fashion. The winning unit will induce the motor action. In case UR does not receive any input from the collector units, which means none of them matches the actual CS prototype, exploration behavior will be executed (forward translation).

The trigger elements are used to be able to chain through LTM sequences (see Fig. 3 for illustration). If the collector unit of segment lq contributed to activating the cell of UR that won

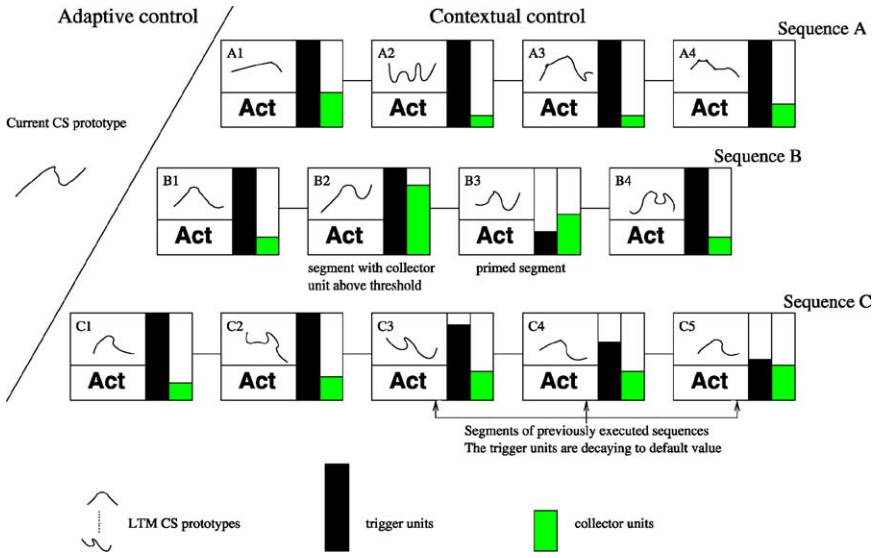

Fig. 3. The matching and chaining of LTM segments: each segment consists of a CS prototype, a related action (Act), a trigger unit (black) and a collector unit (gray). The current CS prototype is matched with those stored in the LTM segments. The matching score and the trigger units give an input to the collector units. If a collector unit is above threshold (B2) it gives an input to population UR (not in this figure) and the trigger unit of the following segment is reduced (B3) to enhance the probability that it will dominate the matching process in the next cycle. The trigger units slowly decay back to their initial value (C3–C5).

the WTA, the value of trigger unit $l+1$ in sequence q will be reduced to a value β ($0 < \beta < 1$). This means that a segment, following a previously effective one, will be given higher priority in future decision making. The activation of the trigger unit of each segment decays to its default value 1 according to:

$$t_{lq}(t+1) = \alpha^t + (1-\alpha^t)t_{lq}(t) \tag{18}$$

2.3. DAC5 as a Bayesian decision maker

By phrasing the foraging tasks performed with the DAC architecture in Bayesian terms we will prove that the DAC architecture will execute exactly those actions that are optimal in a Bayesian sense. Hence, we will define the DAC equivalents of the central components of a Bayesian analysis of the foraging task: goals, actions, hypotheses, observations, experience, prior probabilities and score function.

(A) *Goal*: the goal g is defined by the task, i.e. finding targets and avoiding collisions.
(B) *Actions*: the set $A = \{a_k\}$ consists of all possible motor actions the robot can execute (UR and CR).
(C) *Hypotheses*: the set $S = \{s^{\pm}_{k,n}\}$ corresponds to the set of "speculations" how a target can be reached or a collision suffered. $s^{+}_{k,n}$ stands for the fact that a target will be found n timesteps after executing action a_k. $s^{-}_{k,n}$ is defined analogously for an upcoming collision.
(D) *Observation*: event r is a representation of what the robot "observes" at a location where a decision should be made. This is a representation internal to the system for which we use the recurrent prediction e generated by the adaptive control structure (Eq. (10)).
(E) *Experience*: acquired knowledge is everything stored in LTM.
(F) $p(s^{\pm}_{k,n})$: the prior probability distribution of the hypotheses $s^{\pm}_{k,n}$ is defined as:

$$p(s^{\pm}_{k,n}) = \begin{cases} C_1 & \text{if this has been experienced} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

"if this has been experienced" means that in the past, n timesteps after executing action a_k, a target (collision resp.) occurred. In terms of the learning model this translates to the presence of a segment in a LTM sequence where the nth to last segment has stored action a_k. C_1 is a value that is constant over this set of experiences and chosen such that $p(s^{\pm}_{k,n})$ is normalized.

(G) $p(r|s^{\pm}_{k,n})$: this conditional probability stands for the probability that the robot observes r, given that a target (collision resp.) will occur after n timesteps by executing a_k. The probability $p(r|s^{\pm}_{k,n})$ is only non-zero if there is a sequence in LTM, where the nth to last segment has stored action a_k. Each LTM segment also contains an observation e. The more similar e is to r, the higher the probability to observe r given $s^{\pm}_{k,n}$. As a measure for this similarity we used the activity of the collector unit b_{lq} (Eq. (16)). Hence, the probability $p(r|s^{\pm}_{k,n})$ becomes

$$p(r|s^{\pm}_{k,n}) = \begin{cases} C_2 b_{lq} & \text{if this has been experienced} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

This probability is only non-zero if there is a sequence in LTM, where the nth to last segment has stored action a_k. C_2 is a normalization constant.

(H) *Score function*: the score function $G_g(s^{\pm}_{k,n}, a_{k'})$ indicates the profit if $s^{\pm}_{k,n}$ is true and $a_{k'}$ is executed. In the random foraging tasks considered here the robot is required to maximize the number of targets found while minimizing the number of collisions suffered. Hence, the score function should be positive if action $a_{k'}$ leads to a target and negative if it leads to a collision. It should be 0 ("neutral") if action $a_{k'}$ is not the same as a_k. In addition, the closer the target (obstacle resp.) the higher the probability of really reaching it, by executing the action. So, an appropriate score function would be:

$$G_g(s^{\pm}_{k,n}, a_{k'}) = \begin{cases} \pm\dfrac{1}{n} & \text{if } k = k' \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Given these definitions the optimal action can be calculated in the following way. The conditional probability $p(s^{\pm}_{k,n}|r)$ is the probability that a target or a collision will occur after n timesteps by executing a_k, given observation r. Combining Eqs. (19) and (20) we get from Bayesian theory (Eq. (1)):

$$p(s^{\pm}_{k,n}|r) = \frac{p(r|s^{\pm}_{k,n})\,p(s^{\pm}_{k,n})}{p(r)} = \begin{cases} \dfrac{C_1 C_2 b_{lq}}{p(r)} & \text{if this has been experienced} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $p(r)$ is the probability to observe r. From Eqs. (21) and (22) we get the expectancy (Eq. (2)):

$$\begin{aligned} \langle g \rangle_{a_{k'}} &= \sum_{s^{\pm}_{k,n} \in S} p(s^{\pm}_{k,n}|r) G_g(s^{\pm}_{k,n}, a_{k'}) \overset{(21)}{=} \sum_{s^{\pm}_{k',n} \in S} p(s^{\pm}_{k',n}|r) G_g(s^{\pm}_{k',n}, a_{k'}) \\ &\overset{(22)}{=} \begin{cases} \sum_{s^{\pm}_{k',n} \in S} \pm \dfrac{C_1 C_2 b_{lq}}{n p(r)} & \text{if this has been experienced} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{23}$$

According to Bayes' principle (Eq. (3)), the optimal action to execute is now the one that maximizes the expectancy (Eq. (23)). Since C_1, C_2 and $p(r)$ are all constant over the set S it is the action that maximizes:

$$\langle g \rangle'_{a_{k'}} = \begin{cases} \sum_{s^{\pm}_{k',n} \in S} \pm \dfrac{b_{lq}}{n} & \text{if this has been experienced} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

Cell k of the UR population is only getting input from segments that have stored the same motor action as represented by k while the LTM equivalent of n is z_{lq}. Therefore, expression (24) becomes:

$$\langle g \rangle'_{a_k} = \begin{cases} \sum_{s^{\pm}_{k,n} \in S} \pm \dfrac{b_{lq}}{z_{lq}} & \text{if } A_{\mathrm{k}} \text{ is stored in segment } l \text{ of sequence } q \text{ is satisfied} \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

It can be seen that the input f_k to cell k of the UR population is exactly the value of the expectancy $\langle g \rangle'_{a_k}$. This means that the WTA process among the cells of the UR units will be won by the unit standing for the action which maximizes this expectancy.

It has been shown that the general purpose learning system of DAC5 is a Bayesian inference machine. It chooses those actions that are optimal in a Bayesian sense (using assumptions (A)–(H)). The WTA mechanism of the motor map selects the action, which maximizes the expectancy $\langle g \rangle_a$ (Eq. (2)). In addition, the model is self-contained in the sense that the prior probabilities are constantly updated through adding new sequences to LTM while the observations are learned at the adaptive control level.

The model developed here, DAC5, is based on DAC3 (see Verschure & Voegtlin, 1998 for a detailed description of DAC3). The only difference between DAC5 and DAC3 is that the latter applies a WTA selection mechanism at the level of LTM on the collector unit activities (Eq. (16)). Hence, for DAC3 the motor action executed is solely defined by the winning LTM segment. Moreover competition of the collector unit activities is not biased by the distance between a segment and the goal state (Eq. (17)). Hence, DAC5 is a simpler model than DAC3. We emphasize that in all other respects DAC3 and DAC5 are identical.

3. Experimental environments

As in previous studies the experiments were performed using both simulated and real robots (Verschure & Voegtlin, 1998). Simulations guarantee repeatability over trials and therefore allow a systematic evaluation of a control structure. Only experiments with a real robot, however, allow the exploration of the robustness and generalizability of a model (Mondada & Verschure, 1993).

3.1. Simulation environment BugWorld

Simulations were performed using the simulation environment BugWorld (Almássy, 1993; Goldstein & Smith, 1991). The simulated spherical robot (Fig. 4) uses three types of sensors: a range finder (CS), collision sensors (US^-) and target sensors (US^+). The configuration of the

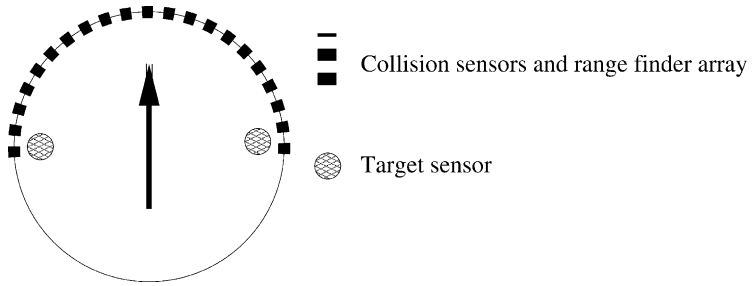

Fig. 4. The simulated robot used in BugWorld. The range finder (black rectangles) consists of 37 sensors distributed over 180° on the front side of the robot. Their angular resolution decreases at the borders (20°) and is maximal at the center (5°). Thirty-seven collision detectors cover the same region as the range finder elements. Two target sensors are located at 90° and −90° from the mid-line of the robot.

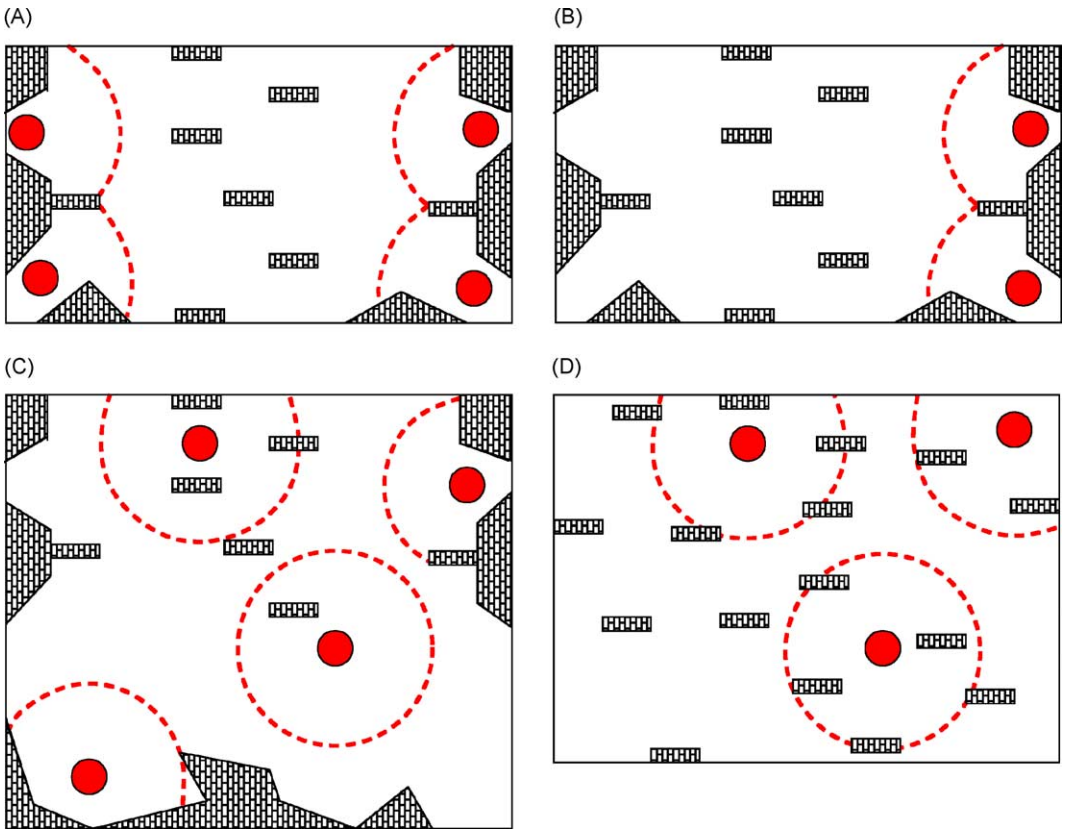

Fig. 5. The four environments used in the simulation experiments, consisting of obstacles (bricks) and targets (solid circles). The region in which the US^+ can be detected by the target sensors is indicated by dashed circles. The soma has the same size as the targets. (A) The standard environment used in previous studies, e.g. Verschure & Voegtlin, 1998. (B) As (A) with non-symmetric target distribution. (C) Large environment with low obstacle density. (D) Environment with high obstacle density.

shape of the robot and the properties of its sensors will be referred to as the *soma*. The soma can execute discrete translational and rotational actions. These actions are coupled together to define behavioral patterns: "exploration," "avoidance," and "approach."

Fig. 5 shows the four different environments used in the simulation experiments. In a secluded space multiple obstacles and targets are placed. Each target can only be detected in a limited region around it. The targets have their own dynamics: in case one of them is touched it is removed. A new target reappears in the same position when another target is found. The distance measure used in the simulations is the body-size of the robot, where its diameter is equal to one unit called *diam*.

3.1.1. The mapping of the sensors and the actions of the soma

The input of the 37 collision sensors (US^-) is mapped to the IS^- population, consisting of 37 neurons. Each sensor is connected to one neuron. The input to each of these neurons is 1 in

case of a collision at the corresponding sensor and 0 otherwise. The IS^+ population consists of two neurons, each connected to one of the target sensors (US^+). The target sensors are active when they are within the target region. If one of the sensors is significantly closer to the target ($\geq$0.25 diam) than the other one, the input to the corresponding neuron is 1 and 0 otherwise. If the two sensors are approximately the same distance to the target ($<$0.25 diam), both neurons in IS^+ are active. The CS population consists of 37 neurons, each corresponding to one of the range finder elements. Their inputs reflect the distance to the next obstacle or wall. The activity, u_j, of CS unit j is defined as:

$$u_j = e^{-\gamma^s s_j} \tag{26}$$

where s_j is the input from range finder element j and γ^s defines the slope of the function. The UR population consists of five neurons each triggering a specific action. The predefined connections between the IS and UR populations define the following responses:

- avoid left (right resp.): one of the neurons in IS^- corresponding to the collision sensors on the right (left resp.) side is active.
- approach left (right resp.): the neuron in IS^+ corresponding to the left (right resp.) target sensor is active and the other one not.
- approach forward: both neurons in IS^+ are active.
- if no neuron in any IS population is active, or the winning neuron in UR is not above threshold, an "explore" action will be triggered: forward translation.

3.2. Khepera-IQR421

Experiments with the microrobot Khepera (K-team, Lausanne, Switzerland) were performed using the distributed simulation environment IQR421 (Verschure, 1997). Khepera (Fig. 6A) is a circular robot with a diameter of 55 mm and a height of 30 mm. The base plate contains the elementary interface to the real-world: effectors and obstacle/light detection. The robot uses two wheels for locomotion. Obstacle and light detection is achieved by eight infra-red send–receive sensors (IR). Six IRs are placed evenly around the front 180° of the robot and two are placed at the back. The angular resolution of each of the IRs is approximately 50°. In addition, the robot is equipped with a color CCD camera (640 × 480 pixels). Khepera was connected to a host computer using a serial port. Only the processes maintaining serial communication, sampling of sensors, and control of effectors were executed locally.

IQR421 supports the study of neural models at different levels of description. It provides a graphical specification language to define, control, and analyze large scale neural simulations using a distributed computing method based on the TCP/IP protocol. In this study five interacting processes were defined; front-end graphics, tracking system, and three simulation and interface processes. Processes communicated synchronously at approximately 10 update cycles/s. The three simulation processes, "Video," "DAC5," and "Khepera," exchange data as indicated by the connections shown in Fig. 6B. "Video" deals with digitizing the video image and simulating the neural system which processes the image. "Video" exchanges the activity of a population of simulated cells reflecting the CS events, with the simulation of the control structure, "DAC5." In addition, "DAC5" receives inputs from populations of simulated cells

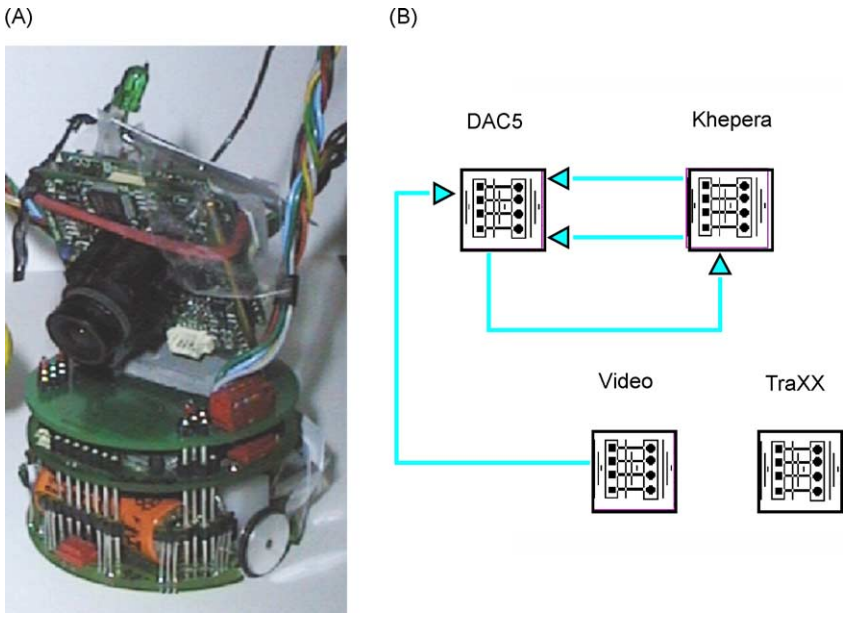

Fig. 6. Real-world experiments. (A) The microrobot Khepera. (B) The four simulation processes defined in IQR421. Each icon stands for one process. Arrows indicate the interactions between these processes.

responding to US events detected by the robot derived from the infra-red sensors. "DAC5" projects the activity of its population expressing URs to "Khepera." "Khepera" in turn interprets its motor map which receives this activity and sends the appropriate commands to the robot. There is also a process "TraXX" which receives input from a tracking camera mounted above the environment and extracts the *x*- and *y*-location of the robot. The environment used in the real-world experiments is depicted in Fig. 7.

3.2.1. The mapping of the sensors and the actions of Khepera

Both $p^{\mathrm{US}-}$ and $p^{\mathrm{US}+}$ (Eq. (5)) were derived from the six frontal IR sensors. On average the IR sensors respond to reflecting surfaces placed up to 5 cm from the sensor. $p^{\mathrm{US}-}$ is defined by thresholding the IR return signal, which gives an approximation of a collision sensor. The raw IR signal was projected on to a population of leaky integrator linear threshold units which rendered $p^{\mathrm{US}-}$. $p^{\mathrm{US}+}$ was derived from the ambient light detected by the six frontal IR sensors in their passive mode. This signal was projected on to a population of leaky integrator linear threshold units. By thresholding with an appropriate value a measure is defined which reflects the presence of a target. The distal sensor, which defines CS events, was provided by the color CCD camera mounted on Khepera. The 480×640 image was compressed to a size of 210×210. Each color channel of the digitized image, using a RGB representation, was pixelized (reduction ratio: 100:1) onto a distinct population of 400 leaky integrators conserving the "retinotopy" of the camera. The CS population consisted of 36 units where each cell reflected a certain combination of particular color channels in sub-regions of the image. For example, a red center surrounded by blue activated a particular cell in the CS population. This was done for

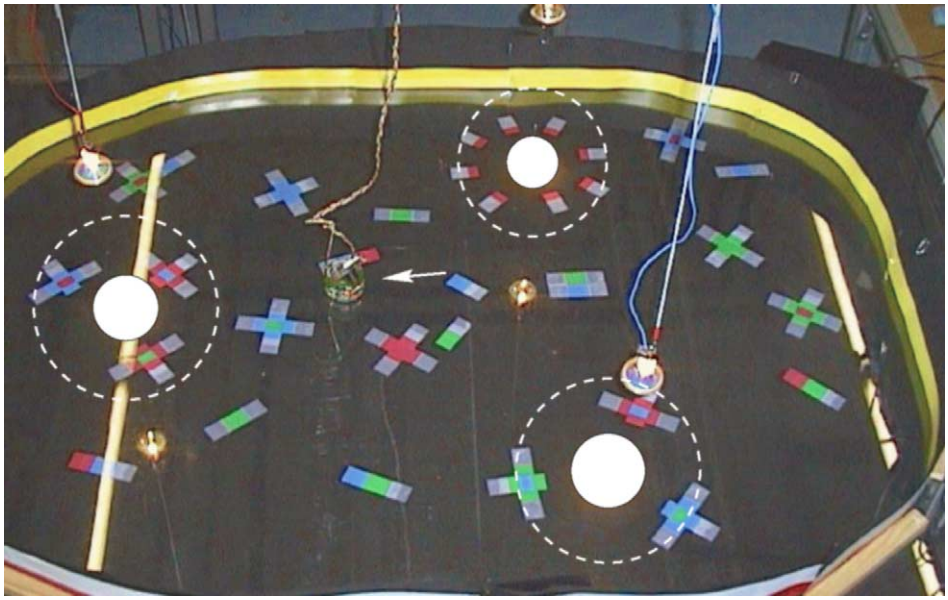

Fig. 7. Environment used for robot experiments: (A) A 1 m × 1.5 m enclosed space contains three target regions (white solid circles) and a number of colored patches (CS). The targets are defined by halogen lamps mounted above the floor that disperse a light gradient. The region in which this gradient can be detected is indicated with white dashed circles. The colored patches consist of combinations of specific surround and center colors (red, green, and blue). The wall surrounding the environment is yellow.

every possible center-surround combination of the colors red, green, blue and yellow. Motor output sent to Khepera was derived from the activity of the UR population. It consisted of eight cells, each of them standing for a certain action (rotational and translational motion). The motor actions belonging to the behaviors "avoid," "approach" and "explore" were defined analogously to the ones for the soma in BugWorld. In this case, as opposed to the simulation, motor activity was continuous. Once initialized the motors only changed their state if another pattern of activity arose (see Verschure & Voegtlin, 1998 for further implementation details).

4. Results

We have pursued a formal approach to prove that the dynamic equations of DAC5 generate actions in a foraging task that are equivalent to those predicted by a Bayesian analysis of the task. This implies that DAC5 is, in the Bayesian sense, an optimal decision making system and obeys the principle of rationality. Hence, DAC5 unifies the perspectives of old and new AI. We observed, however, that in order to make optimal use of the available information, i.e. to achieve the goals of optimizing the number of targets found while minimizing the number of collisions suffered, the DAC architecture needed to be adapted. In order to satisfy the research method of new AI we need to assess whether DAC5 also generalizes to the real-world. Hence, we will assess whether this model shows enhanced performance in random foraging. We performed a

comparative study, using both simulated and real robots, where we evaluated the performance of DAC5, its non-rational predecessor DAC3 and a control condition where the contextual control layer is disabled, called DAC2. As a first qualitative description of the differences between these three models sample trajectories of simulated and real robots are presented. Subsequently a quantitative analysis of the performance is presented.

4.1. Qualitative analysis: BugWorld

In the simulated robot experiments the agent must learn to use the distance profiles provided by its range finders (CS) to efficiently navigate in a secluded space. The knowledge this agent will use to optimize its task, range finder profiles and their relationship with collisions and targets, is not *a priori* defined. The predefined information about targets and collisions is only available through proximal sensors, i.e. these events can only be detected when the agent is in their vicinity. Hence, any form of planning must rely on distal sensor information, i.e. the range finder profiles.

Fig. 8 shows the trajectories of representative examples of DAC2 and DAC5 in the four environments considered (Fig. 5). At the start of each trial the weights of the connections between the CS and IS populations were set to 0 and LTM was empty. The starting position and orientation of the soma were chosen randomly. Each trial lasted 20,000 time steps. The trajectories displayed show the positions visited by the soma during the last 2,000 time steps of a trial.

In the condition where contextual control is disabled, DAC2, the robot visits practically the whole environment, independent of the target distribution. This agent, however, does successfully avoid collisions with walls and obstacles by triggering the avoidance response shortly before a wall or obstacle is reached. This is due to learning at the level of adaptive control that allows the agent to predict upcoming collisions on the basis of its range finder readings. In contrast to this non-rational agent DAC5 shows a much higher specificity in its behavior, repeating similar trajectories. Paths “discovered” earlier in the trial are followed with a higher probability than others, i.e. Fig. 8B, D, F and H. Moreover, the trajectories are more structured around the distribution of the targets in the environment. For instance, in environment B (Fig. 8D) DAC5 remains in the vicinity of the two targets with high probability. In contrast DAC2 does not adjust its behavior to this particular target distribution and shows a practically identical trajectory to that seen in environment A (Fig. 8C).

4.2. Qualitative analysis: real-world

Experiments with the Khepera microrobot have been performed using the environment depicted in Fig. 7. In these experiments the agent has to learn to make use of the colored patches on the floor and the wall of the arena to locate the light sources. Also in this case collisions and light are only detected with the IR sensors (US) when the robot touches an obstacle or is close to the center of the projected light. Hence, for planning the robot must make use of the distal information provided by the colored patches sensed through the CCD camera (CS).

In these experiments the initial conditions were the same as those of the simulation experiments. The trajectories of the Khepera robot (Fig. 9) show the same properties as observed in

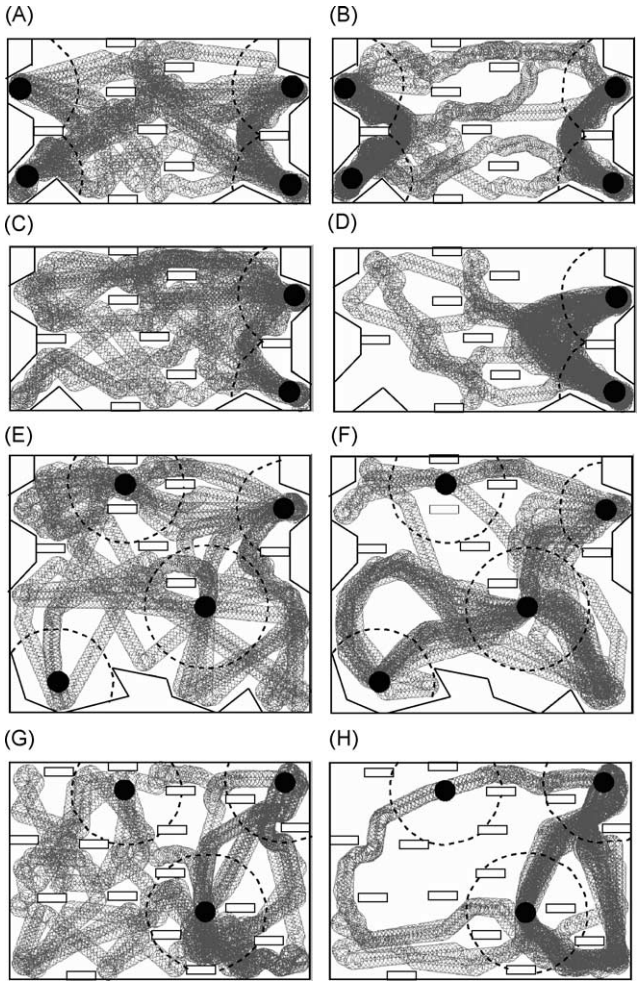

Fig. 8. Simulation experiments: example trajectories in the four different environments depicted in Fig. 5 for DAC2 (left column) and DAC5 (right column). (A and B) Environment A. (C and D) Environment B. (E and F) Environment C. (G and H) Environment D.

the simulation studies. Both models successfully locate the targets and avoid collisions. In the disabled condition the robot again visits most regions in the environment (Fig. 9A). A slightly denser distribution can be observed inside target regions due to the presence of the US^+. In contrast, DAC5 structures its behavior around a small number of prototypical trajectories, only visiting a part of the environment (Fig. 9B).

To assess the impact of learning on performance, recall tests were performed: after 20,000 timesteps the lights (targets) were switched off (Fig. 10). In this case only the colored patches can provide cues to locate the target areas. DAC2 displays highly variable behavior and target regions appear to be visited with a low probability (Fig. 10A). DAC5 shows highly structured behavior organized along the colored patches in the environment, visiting the target regions with a high probability (Fig. 10B).

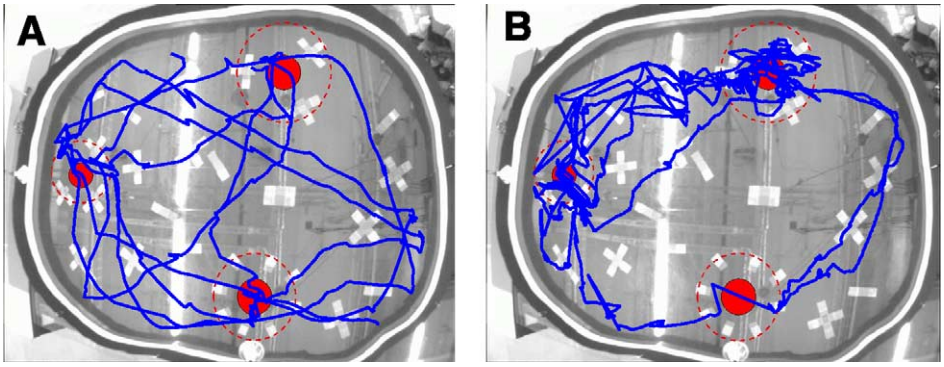

Fig. 9. Example trajectories of the microrobot for DAC2 (A) and DAC5 (B) during acquisition trials. The total length of the acquisition trials was 20,000 time steps. The trajectories shown depict the positions visited by the robot between time steps 15,000 and 18,000. This corresponds to about 25–30 min after the start of the trial.

In order to demonstrate that the behavior of DAC5 observed in the recall test is due to its ability to use its contextual control layer, we determined at what positions actions were triggered due to LTM activity (Fig. 11). LTM dominates the behavior of DAC5 at very distinct points in the environment and is mostly active when a colored patch is in the visual field of the robot. Moreover, most rotations occurred shortly after LTM was active.

In summary, these results show that DAC5 uses its abilities for contextual learning to optimize its trajectories in random foraging tasks in both simulated and real robots in a variety of environments. As a result, it performs more successfully than a system where contextual control is disabled, DAC2.

4.3. Performance quantification

So far we have used a qualitative approach to assess the differences in performance between the rational agent, DAC5, and a control condition where the contextual control layer is disabled,

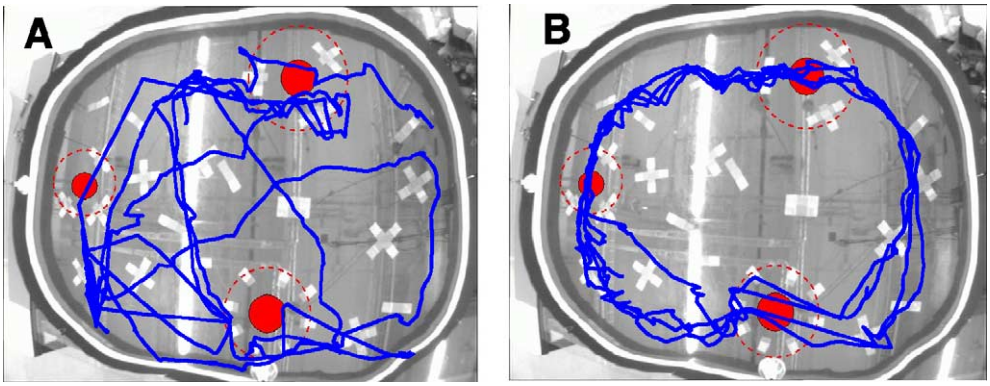

Fig. 10. Example trajectories of the microrobot for DAC2 (A) and DAC5 (B) during recall tests. After 20,000 time steps of learning the targets were switched off. Trajectories shown show the positions visited between time steps 20,000 and 23,000 (from about 33 to 38 min after the start of the trial).

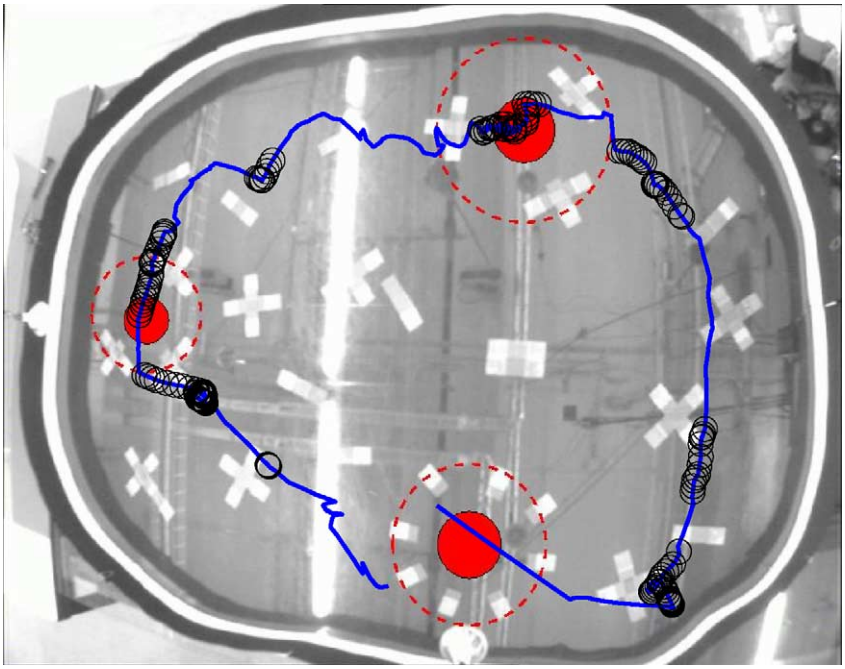

Fig. 11. Usage of LTM by DAC5 in a recall test. Locations in the environment where LTM controlled the action selection are indicated with circles. The trajectory shown was followed in a clockwise direction and generated between time steps 21,400 and 22,000.

DAC2. Although such an approach does provide a useful first order description of a behaving system, it cannot be taken as conclusive evidence. Hence, in order to put our conclusion on a more solid foundation we performed a more detailed quantitative analysis. We evaluated 200 exemplars per condition in the four simulation environments (Fig. 5) in trials lasting 20,000 time steps. A total of 2,400 simulated robots were evaluated where the initial positions and orientations of the robots in the environment were randomized. In this analysis we considered DAC5, its predecessor DAC3, and again a condition where the contextual control layer is disabled, DAC2.

Fig. 12A–D show the performance of the different models in the four environments in terms of the number of targets found and collisions suffered per traveled distance (Fig. 5). Although the environments have quite different characteristics in terms of the amount of targets and obstacles, the performance of the different models evolve, relative to each other, in a similar fashion. In all cases we observe that both performance measures in all four environments tend to converge. Hence, in all cases the learning systems investigated, on the average, developed stable behavior. Initially in all four environments the high collision rate decreases rapidly. This is due to the construction of associations between the CS and US by the adaptive control layer (see also Verschure & Voegtlin, 1998). After reaching the confidence threshold (Eq. (13)) DAC3 and DAC5 activate their contextual control layer around time step 4,000. After this point in time the three models start to perform very differently. Through the use of sequential representations, DAC3 improves its obstacle avoidance behavior compared to DAC2. DAC5,

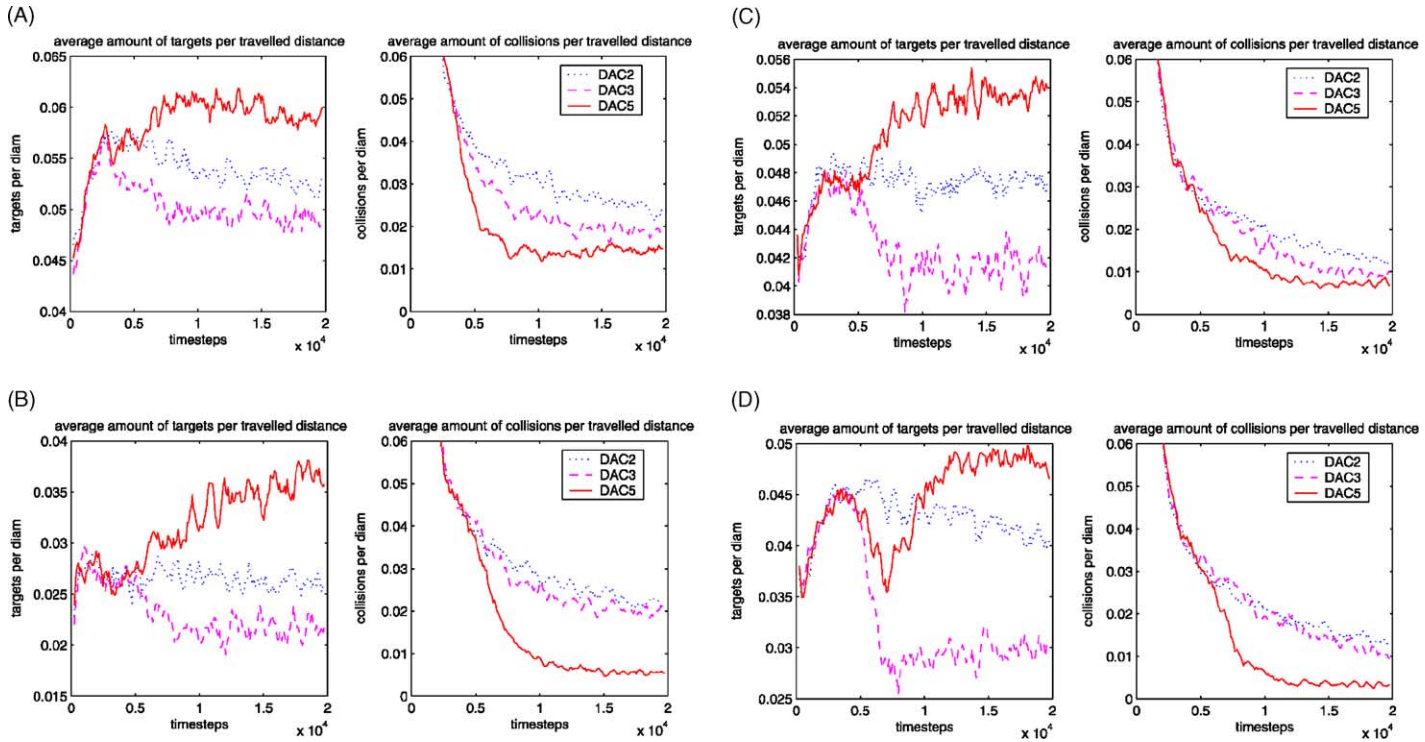

Fig. 12. Performance comparison of DAC2, DAC3, and DAC5 in the different environments depicted in Fig. 5. For each condition 200 exemplars were evaluated. The initial position and orientation of the robots in the environment were randomized. Performance was quantified as the average number of collisions suffered (right column) and targets found (left column). Averages were calculated over all 200 exemplars per condition every 100 time steps for a 500 time step sliding time window. (A and B) Environment A. (C and D) Environment B. (E and F) Environment C. (G and H) Environment D.

however, further reduces its average number of collisions compared to DAC3. The average amount of targets found in case of DAC3 is lower than for DAC2. While also on this measure DAC5 clearly shows better performance. For environment A we observe that the average number of targets found per traveled distance converges to about 0.05 for DAC2 and DAC3 while DAC5 shows an higher average of 0.06. DAC2 converges to the highest average number of collisions in this environment, about 0.025, DAC3 gives a value of 0.02, while DAC5 displays the lowest collision ratio of 0.018.

DAC5 performs better than the other two control models in all four environments, both in finding targets and avoiding collisions. These results demonstrate that the changes made to our DAC architecture to satisfy the principle of rationality have rendered a control structure that also displays superior performance in real-world tasks as compared to its non-rational predecessor and a control condition where the contextual layer was disabled.

5. Discussion

We have shown that the perspectives of traditional and new AI can be unified. We have redefined the knowledge level description of a complex task, including its principle of rationality, in terms of a Bayesian analysis. Subsequently we have proven that a robot based architecture of learning and problem solving generates the same actions as those predicted by the Bayesian analysis. In addition, we have shown that this architecture shows robust performance in random foraging in different environments for both simulated and real robots. DAC5 is a self-contained learning system that demonstrates how problem solving and behavioral control can be understood in strictly bottom-up terms. Using only prewired reflexes a reactive control layer is equipped with a minimal behavioral competence to deal with its environment. The adaptive control layer associates inputs from distal sensors with these reflexes and constructs representations (expectancies) of sensory stimuli. At the level of contextual control these representations are used to form more complex representations in order to express relationships of sensory and motor events over time. It has been shown that the contextual control structure uses these representations in an optimal Bayesian way to achieve its goals in simulated and real foraging tasks. The learning model is self-contained in the sense that the prior and conditional probabilities are acquired through interaction with the world and are continuously updated in relation to the experience of the agent, by changes in the classification of sensory events at the adaptive layer or the formation of new sequences by the contextual layer. A key difference between DAC5 and traditional rational systems is that the former becomes rational due to its continuous interaction with the world while the latter are rational as a result of prior specification. The “symbols” DAC5 is integrating in its optimal decision making are acquired and not predefined. An important consequence of this is that where the rationality of traditional systems is bounded by the logical closure of their predefined world models, that of the DAC architecture is bounded by the complexity of the real-world in which it operates and the direct interfaces it possesses. Moreover, the predefined rules that give rise to global rational behavior are themselves non-rational and only based on local interactions, e.g. local learning at the level of adaptive control, WTA interactions at the level of a motor map or nearest-neighbor chaining in LTM. Moreover, the principles included in the model are based on an analysis of the learning paradigms of classical

and operant conditioning. We have successfully generalized some of these principles towards the neuronal substrate (Hofstoetter et al., 2002; Sanchez-Montanes et al., 2000, 2002). Hence, the DAC architecture described here unifies three key domains. First, it unifies old and new AI by showing that rationality can emerge in a real-world agent out of the combination of acquired "symbols" with their integration in decision making solely based on local rules. Second, it unifies rationality with the principles underlying classical and operant conditioning.

A first challenge to the unification demonstrated here is that a Bayesian framework might not sufficiently capture the knowledge level. This claim can only be evaluated in terms of the definitions of a knowledge level analysis. In a weak form one could equate the knowledge level with an intentional stance towards intelligent behavior dissecting overall functionality in terms of knowledge, goals and actions (Dennett, 1988). In this view many approaches can be taken to perform a knowledge level analysis. Key to the knowledge level, however, is the principle of rationality which imposes a strong constraint on how an intelligent system uses its knowledge. It is exactly this element that is directly captured in a Bayesian analysis. In addition, Newell defines intelligence as the degree to which a system is able to approximate the knowledge level. In this case intelligence is defined as the ability to bring all available knowledge to bear to achieve goals (Newell, 1990). This property is another defining feature of the Bayesian framework. Hence, the choice to operationally define the knowledge level in Bayesian terms is not arbitrary. The Bayesian framework captures the central elements of knowledge level descriptions as elaborated by Newell. Moreover, expressing the knowledge level in Bayesian terms places this functional approach on a solid experimental foundation in both psychology and neuroscience.

A second issue pertains to the relationship between the Bayesian knowledge level account and the DAC5 architecture that ultimately controls the behavior of the robot. We have shown that the DAC5 architecture not only generates the same actions as those predicted by a Bayesian analysis but that it actually directly implements key components of the Bayesian analysis. For instance, the dynamics of the collector units of LTM (Eq. (16)) directly reflect the conditional probabilities on which a Bayesian analysis is based. This could be seen as an argument in favor of a strictly functionalist perspective where the new AI component of the approach presented here only shows how a functional analysis of intelligent behavior can be implemented (Fodor & Pylyshyn, 1988). Our results, however, demonstrate that such an interpretation is not valid. The DAC5 architecture comprises multiple components, but only a subset of these, in particular its contextual control structure, give rise to its rational behavior. In isolation, however, this set of mechanisms would not give rise to any behavior. At start-up the system is devoid of any hypothesis on how it can plan its behavior. These hypothesis are acquired and updated, at the level of adaptive and contextual control, by virtue of a continuous interaction with the real-world. Hence, this system can only display rational behavior by being in the real-world. The embodiment of our system also has consequences for its knowledge level description. For instance, by virtue of being in the world DAC5 does not need to *a priori* represent all possible states of its world. It will only acquire a small subset of all possible states that is relevant to its behavior. Moreover, as opposed to traditional systems that require complex operations to assess goal achievement, DAC5 only requires a minimal set of mechanisms to evaluate its ability to achieve its goals, i.e. the detection of targets or collisions. Hence, the unification presented here not only demonstrates that situated agents can be rational, it also constraints a further knowledge level analysis of the task under consideration. This is in accord with earlier criticism of the

knowledge level, where it was argued that this perspective is inherently under-constrained and needs to be placed in the context of physical systems interacting with the real-world (Clancey, 1989b). Another consideration is that functionalism, as also expressed in traditional AI models such as SOAR (Newell, 1990) and ACT (Anderson, 1983), succeeded in describing complex cognitive processes but was not able to deal with the fundamental symbol grounding and frame problem. The unification presented here shows that these cognitive processes can still be effectively described and explained without paying this price. This is relevant to cognitive science since it argues for an integrated approach towards the study of mind, brain, and behavior as opposed to one that solely relies on the study of a disembodied mind. We consider it a great advantage to show how traditional and new AI can be unified in order to resolve the apparent conflict between the rationalist and empiricist traditions on which they rest.

Any bottom-up approach towards learning has to justify the *a priori* design elements of the system. For DAC5 this pertains to the predefined properties of sensory-motor control and its learning mechanisms. At start-up the only semantics available to the system are those defined at the reactive control layer. These include only events detected by proximal sensors that reflect the immediate presence of targets or collisions. The sensory representations used in decision making, defined by e (Eq. (10)), are not defined *a priori* but are acquired by the adaptive layer using low complexity signals, i.e. collisions and targets, provided by the reactive layer. The observation r, used in the Bayesian analysis, could also have been defined using other measures than e that do not depend on learning, i.e. the uninterpreted immediate state of the distal sensor. However, we observed that using e gives markedly better results (data not shown). The contextual control layer of DAC5, that provides the substrate for optimal decision making, is *a priori* devoid of any hypothesis on how states of the world relate to goal-oriented behavior. Moreover, the rule for the storage of observations and actions were kept minimal, i.e. all CS events and associated actions were stored in STM. Also the retention of a STM sequence in LTM depended on minimal assumptions, i.e. the occurrence of a collision or target event.

Our results show that DAC5 is able to achieve high performance in random foraging tasks compared to our controls. Even after removing the target stimulus in recall tests, the target regions were reliably revisited. Moreover, DAC5 organizes its behavior differently to the adaptive control structure (DAC2) converging on a smaller number of trajectories between target locations. We demonstrated that this difference is due to the use of the contextual control layer. One could argue, however, that the trajectories displayed by DAC5 (Fig. 8) are not the shortest path between targets and hence not optimal from an objective point of view. However, they can be considered optimal, given the system's incomplete knowledge of the world akin to Simon's notion of bounded rationality (Simon, 1969). This, however, raises the important question how an inductive system can optimize its problem solving behavior beyond its direct experience, or the exploration–exploitation dilemma (Kaelbling, Littman, & Moore, 1996). DAC5 only has a very limited ability to explore its physical problem space, translational action, and will not critically reevaluate the plans for behavioral control it has acquired. We speculate that the combination of more variable exploration behavior with the ability to override predictions generated by the contextual control structure will improve the problem solving abilities of our system without violating its strict bottom-up design principles.

In (Saksida, Raymond, & Touretzky, 1997; Touretzky & Saksida, 1997) a robot based model of operant conditioning is presented as an extension of reinforcement learning. The model

replicates some fundamental phenomena associated with instrumental learning. However, it needs a human to provide reinforcement to the system. In DAC5 the reinforcement signal that induces the retention of a STM sequence in LTM is triggered autonomously. Bayesian approaches have been applied to robots, for instance, in obstacle avoidance tasks (Hu & Brady, 1994) and the processing of noisy sensor data in the context of path planning (Kristensen, 1997). Contrary to DAC5, however, in these cases the priors included in the model are defined *a priori* by the designers of the system. Although such an approach might be advantageous from an engineering perspective it does imply that these solutions are not self-contained. An approach closest to the model described here is applied to landmark learning in the context of the robot localization problem (Thrun, 1998). In this case a model, called BaLL, is presented that enables a mobile robot to learn what features are best suited for localization. This model, however, solely focuses on assessing the optimal use of sensory data to reduce the error in the estimate of a robot's location and does not address the issue of goal-oriented behavior. Moreover, like traditional approaches, BaLL aims at developing a global representation of a task, i.e. an environment, allowing the robot to "know" its location at any one point in time. The model presented here proposes that these acquired representations can be limited to a number of prototypical behavioral sequences that support goal-oriented behavior.

DAC5 was restricted in the sense that targets are treated as "positive" and collisions as "negative" events (Eq. (17)). Hence, the system could only learn to maximize or minimize a fixed set of goal states. Biological systems do explicitly represent the quality of behavioral states (Schultz & Dickinson, 2000), however, they are not assigned on a fixed *a priori* basis to specific events but again subject to learning (Hollerman & Schultz, 1998). An example of this can be found in so-called secondary conditioning (Rescorla, 1980). In this case after an animal has been trained to respond to a CS, e.g. a tone, this stimulus itself can become a reinforcer, i.e. signaling a potential goal state. The current architecture already explicitly represents the elements of secondary conditioning at the level of the IS populations of the adaptive control layer. We believe that adding the ability to learn goal states would allow our DAC5 architecture to generalize to a wider range of tasks, e.g. solving impasse situations. This might be providing a way for the learning model to define its tasks itself through the dynamics of the IS populations. This generalization would bring the embodied problem solving system presented here closer to the goal of general intelligence, where "within some broad limits anything can become a task" (Newell, 1980).

Acknowledgments

Part of this project was supported by the Swiss National Science Foundation—SPP and the Volkswagen Foundation.

References

Almássy, N. (1993). *BugWorld: A distributed environment for the development of control architectures in multi-agent worlds* (Technical report 93.32). Department of Computer Science, University Zurich.

Anastasia, T. J., Patton, P. E., & Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, *12*(5), 1165–1187.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge, MA: MIT Press.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Transactions of the Royal Society*, *53*, 370–418.

Brooks, R. (1991a). New approaches to robotics. *Science*, *253*, 1227–1232.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Burgess, A. (1985). Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. *Journal of Optical Society of America A*, *2*, 1498–1507.

Chandrasekaran, B. (1994). Understanding control at the knowledge level. In *Working notes AAAI fall symposium on control of the physical world by intelligent agents* (pp. 19–26). AAAI.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*(2), 57–65.

Clancey, W. J. (1989a). The frame of reference problem in cognitive modeling. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 107–114). Hillsdale, NJ: Erlbaum.

Clancey, W. J. (1989b). The knowledge level reinterpreted: Modeling how systems interact. *Machine Learning*, *4*, 285–291.

Clancey, W. J. (1996). *Situated cognition: On human knowledge and computer representations*. Cambridge: Cambridge University Press.

Dennett, D. C. (1988). *The intentional stance*. Cambridge, MA: Bradford Books/MIT.

El-Gamal, M. A., & Grether, D. M. (1995). Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association*, *90*, 1137–1145.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture, a critical analysis. *Cognition*, *28*, 3–71.

Fong, C., & McCabe, K. (1999). Are decisions under risk malleable? *Proceedings of the National Academy of Sciences of the United States of America*, *96*(19), 10927–11132.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Goldstein, L., & Smith, K. (1991). *BugWorld a distributed environment for the study of multi-agent learning algorithms* (Technical report). Department of Computer Science, UCSC.

Haddawy, P. (1999). An overview of some recent developments in Bayesian problem-solving techniques. *AI Magazine*, *20*, 11–19.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act*. Cambridge MA: MIT press.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behaviour*, *13*, 243–266.

Hofstoetter, C., Mintz, M., & Verschure, P. F. M. J. (2002). The cerebellum in action: A simulation and robotics study. *European Journal of Neuroscience*, *16*, 1361–1376.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

Hu, H., & Brady, M. (1994). A Bayesian approach to real-time obstacle avoidance for a mobile robot. *Autonomous Robots*, *1*, 69–92.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–385.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, *82*, 35–45.

Knill, D. K., & Richards, W. (1996). *Perception on Bayesian inference*. Cambridge, MA: Cambridge University Press.

Koechlin, E., Anton, J. L., & Burnod, Y. (1999). Bayesian inference in populations of cortical neurons: A model of motion integration and segmentation in area MT. *Biological Cybernetics*, *80*, 25–44.

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, *275*, 27–31.

Kristensen, S. (1997). Sensor planning with Bayesian decision theory. *Robotics and Autonomous Systems*, *19*, 273–286.

Laird, J. E., & Rosenbloom, P. (1996). The evolution of the SOAR cognitive architecture. In David M. Steier & Tom M. Mitchell (Eds.), *Mind matters: A tribute to Allen Newell* (pp. 1–50). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mackintosh, N. J. (1974). *The psychology of animal learning*. New York: Academic Press.

Massaro, D. W. (1997). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*, 225–252.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence, 4*, 463–502.

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

Mellers, B. A., Schwartz, A., & Cooke, A. D. (1998). Judgment and decision making. *Annual Review in Psychology, 49*(1), 447–477.

Mondada, F., & Verschure, P. F. M. J. (1993). Modeling system–environment interaction: The complementary roles of simulations and real world artifacts. In J. L. Deneubourg, H. Bersini, S. Goss, G. Nicolis, & R. Dagonnier (Eds.), *Proceedings of the Second European Conference on Artificial Life* (pp. 808–817). Cambridge, MA: MIT Press.

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science, 257*, 1357–1363.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135–183.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87–127.

Newell, A. (1990). *Unified theories of cognition*. Cambridge MA: Harvard University Press.

Newell, A. (1992). Precis of unified theories of cognition. *Behavioral and Brain Sciences, 15*, 425–492.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). A general problem solving program for a computer. *Computers and Automation, 8*, 10–16.

Newell, A., & Simon, H. A. (1963). GPS, a program that simulates human thought. In J. Feldman & E. A. Feigenbaum (Eds.), *Computers and thought*. New York: Mc Graw-Hill.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communication of the ACM, 19*, 113–126.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Pfeifer, R. (1995). Cognition: Perspectives from autonomous agents. *Robotics and Autonomous, 15*, 47–70.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*, 233–238.

Porrill, J., Frisby, J. P., Adams, W. J., & Buckley, D. (1999). Robust and optimal use of information in stereo vision. *Nature, 397*, 63–66.

Rao, R. (1999). An optimal estimation approach to visual perception and learning. *Vision Research, 39*(11), 1963–1989.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87.

Rescorla, R. A. (1980). *Pavlovian second-order conditioning: Studies in associate learning*. Hillsdale: Erlbaum.

Roberts, W. A. (1992). Foraging by rats on a radial maze: Learning, memory, and decision rules. In I. Gormezano & E. A. Wassermann (Eds.), *Learning and memory: The behavioural and biological substrates* (pp. 7–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Saksida, L. M., Raymond, S. M., & Touretzky, D. S. (1997). Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems, 22*, 231–249.

Sanchez-Montanes, M. A., König, P., & Verschure, P. F. M. J. (2002). Learning sensory maps with real-world stimuli in real time using a biophysically realistic learning rule. *IEEE: Transactions on Neural Networks, 13*, 619–632.

Sanchez-Montanes, M. A., Verschure, P. F. M. J., & König, P. (2000). Local and global gating of plasticity. *Neural Computation, 12*, 519–529.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Reviews in Neuroscience, 23*, 473–500.

Searle, J. (1982). Minds brains and programs. *Behavioral and Brain Sciences, 3*, 417–424.

Simon, H. A. (1969). *Sciences of the artificial*. Cambridge, MA: MIT Press.

Suchman, L. A. (1987). *Plans and situated actions*. Cambridge: Cambridge University Press.

Thorndike, E. L. (1911). *Animal intelligence*. New York: Macmillan.

Thrun, S. (1998). Bayesian landmark learning for mobile robot localization. *Machine Learning, 33*, 41–76.

Touretzky, D. S., & Saksida, L. M. (1997). Operant conditioning in Skinnerbots. *Adaptive Behavior, 5*, 219–247.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Verschure, P. F. M. J. (1993). Formal minds and biological brains. *IEEE Expert, 8*(5), 66–75.

Verschure, P. F. M. J. (1997). *Xmorph: A software tool for the synthesis and analysis of neural systems* (Technical report). Institute of Neuroinformatics, ETH-UZ.

Verschure, P. F. M. J., (1998). Synthetic epistemology: The acquisition, retention, and expression of knowledge in natural and synthetic systems. In *Proceedings World Conference on Computational Intelligence 1998. Anchorage* (pp. 147–153). IEEE.

Verschure, P. F. M. J. (2000). The cognitive development of an autonomous behaving artifact: The self-organization of categorization, sequencing, and chunking. In H. Ritter, H. Cruse, & J. Dean (Eds.), *Prerational intelligence: Adaptive behavior and intelligent systems without symbols and logic* (Vol. 2, pp. 469–488). Dordrecht: Kluwer Academic Publishers.

Verschure, P. F. M. J., Kröse, B., & Pfeifer, R. (1992). Distributed adaptive control: The self-organization of structured behavior. *Robotics and Autonomous Systems, 9*, 181–196.

Verschure, P. F. M. J., & Pfeifer, R. (1992). Categorization, representations, and the dynamics of system-environment interaction: A case study in autonomous systems. In J. A. Meyer, H. Roitblat, & S. Wilson (Eds.), *From animals to animals: Proceedings of the Second International Conference on Simulation of Adaptive Behavior, Hawaii, Honolulu* (pp. 210–217). Cambridge, MA: MIT press.

Verschure, P. F. M. J., & Voegtlin, T. (1998). A bottom-up approach towards the acquisition, retention, and expression of sequential representations: Distributed adaptive control III. *Neural Networks, 11*, 1531–1549.

Verschure, P. F. M. J., Wray, J., Sporns, O., Tononi, G., & Edelman, G. M. (1995). Multilevel analysis of classical conditioning in a behaving real world artifact. *Robotics and Autonomous Systems, 16*, 247–265.

Vinkhuyzen, R. E., & Verschure, P. F. M. J. (1994). The legacy of Allen Newell: Unified theories of cognition. *American Journal of Psychology, 107*(3), 454–464.

Voegtlin, T., & Verschure, P. F. M. J. (1999). What can robots tell us about brains? A synthetic approach towards the study of learning and problem solving. *Reviews in the Neurosciences, 10*(3–4), 291–310.

Weiss, Y., & Adelson, E. H. (1998). *Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision*. A.I. Memo 1624, Dept. of Brain and Cognitive Sciences, MIT.