Submitted to: "ICCM2006 – The 7th International Conference on Cognitive Modeling"

Integrating Reinforcement-Learning and Accumulator Models to Study Decision Making and Reaching Behaviours

Dimitri Ognibene (dimitri.ognibene@istc.cnr.it)
Francesco Mannella (francesco.mannella@istc.cnr.it)
Giovanni Pezzulo (giovanni.pezzulo@istc.cnr.it)
Gianluca Baldassarre (gianluca.baldassarre@istc.cnr.it)
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche,
Via San Martino Della Battaglia 44, 00185 Roma, Italy

Abstract

This paper presents a model of organisms' sensorimotor system based on the idea that complex behaviors are build on the basis of repertoires of sensorimotor primitives organised around specific goals (in this case, arm postures). The architecture of the model incorporates an actor-critic reinforcement learning system, enhanced with an "accumulator model" for decision making, capable of selecting sensorimotor primitives to accomplish a discrimination reaching task that has been used in physiological studies of monkeys' premotor cortex. The results show that the proposed architecture is a first important step towards the construction of a biologically-sound integrated model of the primitive-based hierarchical organization of organisms' sensorimotor systems.

Introduction

The goal of this paper is to present a model of how the sensory-motor system of organisms might be organized on the basis of a repertoire of sensory-motor primitives (Arbib, 1981) that are suitably "assembled" to produce more complex behaviors. This issue is important both for understanding organisms and for building artificial intelligent systems (Schaal, 1999; Rainer & Tani, 2004).

The first specific goal that the model pursues is to understand how organisms build repertoires of sensory-motor primitives on the basis of experience. In particular, the paper will focus on the acquisition of primitives related to the generation of arm's postures in space. It has been shown that neural systems of organisms of various species, from insects and amphibians to mammals and humans is organized around "macro-actions" that when triggered tend to accomplish a change of the environment, or of own body, that represnts a whole "goal". For example Giszter, Mussa-Ivaldi & Bizzi (1993) showed how if some regions of the spinal cord of a frog are stimulated electrically, the limbs of them tend to perform movements that bring them to a given resting point independently of the starting position. Remarkably, there seem to be a relatively small number of these "macro-actions", whose origin is likely filogenetic, encoded in the spinal cord. Similarly, Graziano, Taylor & Moore (2002) showed that if the premotor cortex of monkeys is stimulated electrically, their arms tend get a given posture in space. In humans, a great part of sensorimotor skills are acquired during the first years of life without direct rewards and on the basis of self-generated experience (von Hofsten, 1982). These skills involve both low level capabilities of getting postures in space or of generating cyclical movements (through "central pattern generators", cf. Swanson, 2005, and Schaal, 1999: these are not tackled here).

A second specific goal of the paper is to study how organisms can assemble the motor primitives to accomplish complex tasks. Increasing evidence is showing that basal ganglia (Kandel, Schwartz & Jessell, 2000), some nuclei that at the base of the forebrain of vertebrates and that receive incoming signals from virtually the whole cortex and send signals to the motor part of it (pre-frontal, premotor and motor cortex) via the thalamus, might play an important role in this (Baldassarre, 2002). First of all, dopaminergic neurons basal ganglia are involved in classical conditioning-like tasks where originally neutral stimuli progressively acquire the role of predictors of *primary rewards* through experience (Shultz, Dayan & Montague, 1997). These processes have been successfully modelled on the basis of actor-critic reinforcement learning architectures (Barto, Sutton & Anderson, 1983). Indeed, many aspects of the "critic" component of this architecture, based on the TD-learning algorithm (Barto & Sutton, 1998), has been shown to mimic very well some aspects of the dopaminergic physiology of the basal ganglia (Houk, Davis & Beiser, 1995). On the other side, the "actor" component of the model, that should mimic the sensorimotor part of the basla ganglia, has been modeled and related to specific known physiological mechanisms of the brain in much less detail. With this respect, the third goal of the model is to start to develop an actor that on one side is more closely related to the known mechanisms operating in the brain, and on the other side is closely integrated into a system that implements the idea of sensorimotor primitives illustrated above. An important solution is inspired by Schall (2001), that has shown that when a monkey has to accomplish an oculomotor saccade on one among some alternative targets, some neurons of the frontal-eye-field (premotor cortex) give place to a sort of "race" in which different (groups of) cells "accumulate evidence" (activation) in favour of the different

options: the first (group of) cell(s) that reaches a threshold triggers a saccade towards the target corresponding to it. This processes have been successfully modelled through "accumulators-based models" (Usher & McClelland, 2001), that are probably the best available biologically-plausible models of decision making. The architecture presented here will use one of these models to accomplish the aforementioned goal of building an enhanced "actor".

The system has been trained and tested by using a "discrimination reaching task" similar to that used by Cisek & Kalaska (2005) to carry out physiological recordings in monkeys' premotor cortex (this task showed to trigger "races" among the premotor neurons similar to those mentioned above). The task is composed of five phases (see bottom part of Figure 4): (1) "center-hold time – cht": the hand of the monkey is positioned at a central starting position of an horizontal plane in front of it, and a green cue circle appears at the centre of a screen set in front of the subject; (2) "spatial cue – sc": a red and a blue cue circles (2 cm of radius) appear on the screen at two opposed positions of eight possible target locations distributed around a circle; (3) "memory – mem": a green cue circle appears again at the centre of the screen; (4) "color cue – cc": a color cue, either red or blue, appears at the centre of the screen: this nonspatial cue signals which of the two memorized color-coded spatial cue locations is the target that the monkey should reach; (5) "go signal – go": 8 green circles appear at all the possible target locations: if the monkey reaches the target position that matches both one of the two spatial cues *and* the color cue, it receives a reward. In the simulations, the first 4 phases last 1 s. each, while the fifth lasts 6 s.

The rest of the paper first presents the architecture of the model, then presents the results obtained with it, and finally concludes illustrating the model's strengths and weaknesses.

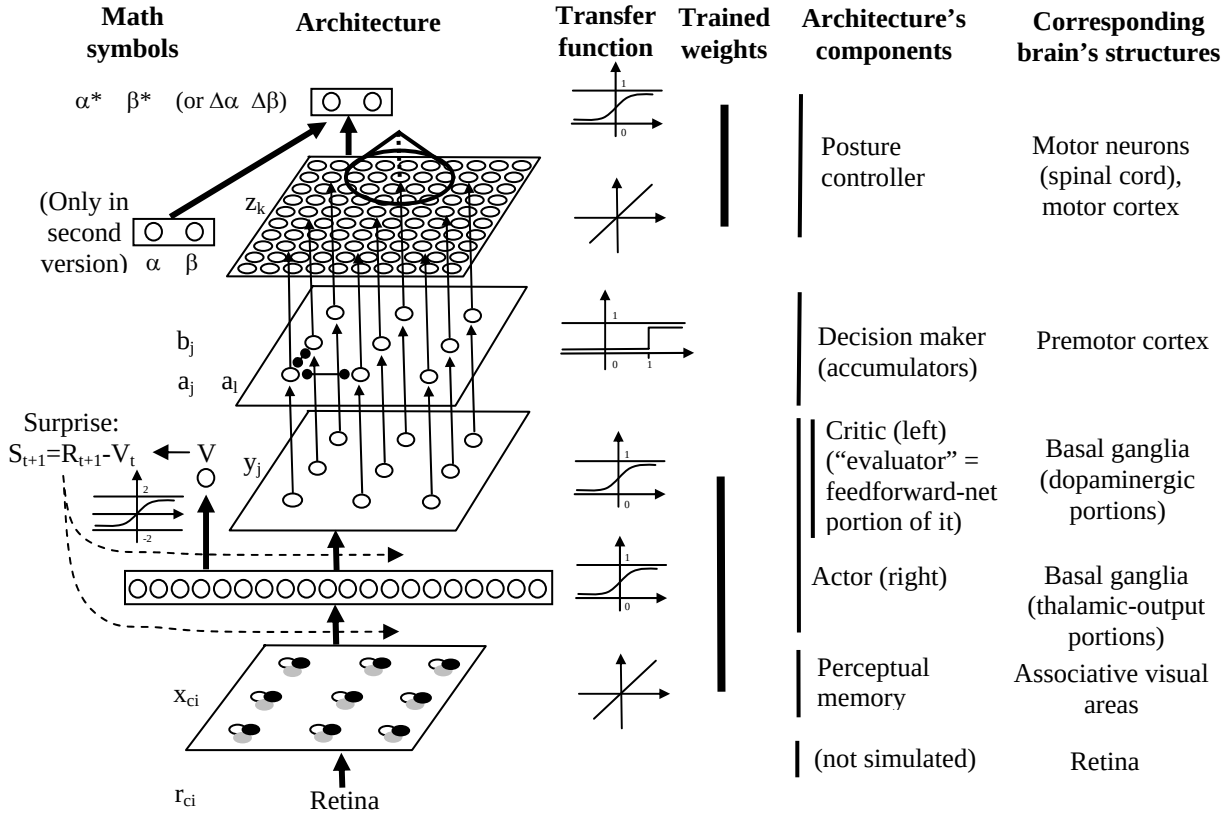

Figure 1: Various aspects of the architecture with the corresponding brain components. Symbols: empty, gray and black small circles (perc. mem.): cells sensitive to green, red and blue; bold arrows: all-to-all connections; arrows: one-to-one connections (weight = 1); dot-head arrows: reciprocal inhibition connections (only few of them have been drawn); dotted arrows: surprise learning signal. The cone within the posture controller map shows an example of one cell's receptive field.

Methods

The "body" of the system is made up by a 2-segments arm that moves on a 2D plane (see Figure 2) plus a retina (not simulated here, see below). The arm's segments measure 20 cm each. Each segment has one degree of freedom: the upper arm can move 180° with respect to the system's "torso" by pivoting on the "shoulder" joint, while the "forearm" can move 180° with respect to the upper arm by pivoting on the "elbow" joint. In this work only simple kinematics of the arm are simulated.

The architecture of the controller of the system (Figure 1) is formed by the following components: perceptual memory, actor-critic, decision maker, and posture controller. The

"life" of the system can be divided in two phases: (1) a "childhood learning phase": the system learns to produce suitable movements that bring the arm to desired postures (encoded in the posture controller's input layer) on the basis of self-generated experience; this learning phase is based on the updating of the posture controller's weights; (2) "adulthood learning phase": the system learns by reinforcement learning to accomplish the discrimination reaching task illustrated in the introduction; this learning phase is based on the updating of the actor-critic's weights. After these two phases, various parts of the system are tested to evaluate their behavior. Now first the architecture and functioning of the system's components are illustrated, and then the two learning phases are explained in detail.

Figure 2: Hand's errors (gray segments) between desired postures (dark dashed arm) and actual postures (black arm), achieved from various initial postures (light dashed arm): (a) before training; (b) after training the controller's first version; (c) after training the controller's second version. (d) a test where the arm has to reach a target posture (dashed arm) in more than one step.

Perceptual memory. This is a map of units x_{ci} ($c \in \{g, r, b\}$; $i \in I$, $|I|$=9) assumed to correspond to associative visual areas. These units receive topological excitatory connections from the retina (not simulated here), are sensitive to different colors (green, red, and blue), and have a leaky-integrator dynamics of the following type:

$$x_{ci,t+1} = \min\left[\left(x_{ci,t} - \tau\left(\kappa x_{ci,t}\right) + r_{ci,t+1}\right), 1\right]$$

where min[.] is a function that guaranties that x_{ci}<1, τ is the integration time step (τ=0.1), κ is the decay coefficient (κ=0.1), and r_{ci} is the color signal from the retina.

Actor-critic. This is a standard feedforward network with 20 hidden units, nine input units (x_{ci}) and ten sigmoid output units: 3x3 units of these, located on a 2D map, are used to select actions's targets (*actor*), and one is used to produce evaluations V_t of perceived states (*evaluator* part of the *critic*: these evaluations are mapped onto [-2, 2] and with learning they tend to become predictors of the reward, see below). The actor-critic network is a neural implementation of a *one-step* actor-critic architecture (Sutton and Barto, 1998). The evaluator's output is used to compute the critic's surprise S_{t+1} signals used to train both the actor and the evaluator (see below) on the basis of the reward R_{t+1}:

$$S_{t+1} = R_{t+1} - V_t$$

Notice that, given this actor-critic architecture and the decaying perceptual memory's units, the system is similar to a reinforcement-learning system that tackles the *time credit assignment problem* on the basis of *eligibility traces* (cf. Sutton and Barto, 1998).

Decision maker. The decision maker is composed of a 2D map of *accumulator* units with activation b_j and activation potential a_j ($j,k \in A$, $|A|$=9: for simplicity, in this research only nine units corresponding to the eight targets, plus the central starting position, were used). Each of these units is activated by the one topologically corresponding actor's unit. The accumulator units have lateral inhibitions and give place to a (noisy) competition that integrates in time, and amplifies, the signals from the actor until one of them reaches a threshold T (T=1), gets an activation b_j=1 and triggers the corresponding goal-posture in the posture controller:

$$\Delta a_j = \tau\left(-\delta\, a_{jt} - \iota \sum_{k \in A, k \neq j} \left[a_k\right] + y_{jt} + \varepsilon \right)$$

$$b_j = 0 \;\; if \;\; a_j < T, \;\; else \;\; b_j = 1$$

where δ is a decay coefficient (δ =0.1), ι is an inhibition coefficient (ι = 0.9), ε is a noise component ranging over [-0.1, +0.1], and τ is the integration step seen previously. The units have a step activation function that returns 1 if the activation potential is above 1, and 0 otherwise. For simplicity, in the experiments presented here the race takes place from the beginning to the end of the "go" phase.

Posture controller. This component has a 2D layer of input-units with an activation z_k ($z \in Z$, $|Z|$=10x10=100). While the component is trained (see below), the input units are activated in [0, 1] on the basis of the (x, y) position of the arm's "hand" on the plane and on the basis of a cone-like activation field having a ray equal to the distance between three units in a row (see Figure 1: during this

training all the 100 units are used; note: the activation of the single units was normalized so that in total it was equal to 1). In contrast, when the component is used to select a target postures each of these units takes as input the activation of the corresponding accumulator unit of the decision maker (for simplicity, during this process only nine units are used). In a second version, the component has two further input units with activation α and β ($\alpha,\beta \in [0, 1]$) that encode the angles of the arm normalized in [0, 1]. The two versions of the posture controller correspond respectively to an "implicit" and an "explicit" modeling of the function played by *fiber-muscle afferents* (sensors located in the muscles, such as the *Golgi tendon-organs*, that return information such as muscles' length to the spinal cord and brain, cf. Shadmehr & Wise, 2005). All the input units are connected all-to-all to 2 sigmoid output units whose activation either encodes the desired angles of the arm, in which case they are denoted as α' and β' ($\alpha',\beta' \in [0, 1]$; first version of the component), or the desired change of them, in which case they are denoted as $\Delta\alpha$ and $\Delta\beta$ ($\Delta\alpha,\Delta\beta \in [0, 1]$; second version of the component).

Table 1: The weights emerged during training of the first version of the posture controller.

	From input units:				
To output units:	**Current α**	**Current β**	**Desired α**	**Desired β**	**Bias**
Angle α	**-5.2820**	-0.0058	**5.5135**	-0.0192	-0.2321
Angle β	-0.0158	**-5.3313**	-0.0239	**5.4660**	-0.1318

Childhood learning phase (posture controller). During this phase the posture controller is trained to perform movements that allow it to reach points on the horizontal plane with its "hand". The target points to reach are encoded as (x, y) Cartesian coordinates on the input-unit 2D map of the controller: one of these targets activates the units of this map within [0, 1] on the basis of their conic receptive field (see above). Training, that mimic self-generated experience, is based on a direct inverse-modeling procedure (cf. Kuperstain, 1988): (1) the current angles of the arm are used to activate with (α, β) the two corresponding input units (only in the second version of the component); (2) a random action, corresponding to a couple ($\Delta\alpha$, $\Delta\beta$), is drawn in terms of variations of the arm's angles within [-10°, +10°] and without violating the limits of the arm's degrees of freedom; (3) the arm performs the movements corresponding to ($\Delta\alpha$, $\Delta\beta$); (4) the new angles (α^*, β^*) of the arm, and position (x^*, y^*) of the hand, are recorded; (5) first variant: an error backpropagation algorithm (Rumelhart et al., 1986; learning rate=0.1) is used to train the posture controller network to associate (x^*, y^*), taken as input, with (α^*, β^*) used as desired output; second variant: an error backpropagation algorithm (learning rate=0.1) is used to train the posture controller network to associate (α, β) and (x^*, y^*), taken as input, with ($\Delta\alpha$, $\Delta\beta$) used as desired output. With this training procedure the posture controller should learn to perform movements that lead it to reach the *desired goal (x^*, y^*) from any initial posture* (note: the first variant of the posture controller is allowed to reach the posture (α', β') that it associates with (x^*, y^*) through "steps" having a maximum size of 10°).

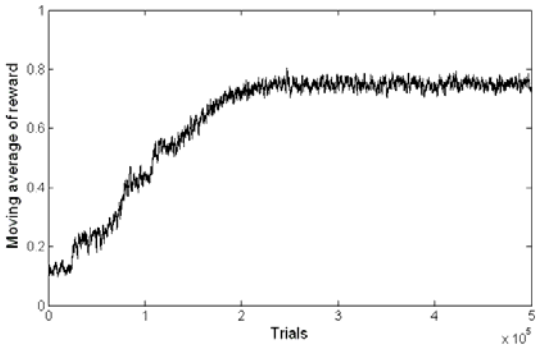

Figure 3: Moving average (1000-step window) of rewards obtained by the system during 500,000 trials of learning.

Adulthood learning phase (actor-critic). During this phase the actor-critic component is trained while the system experiences 500,000 trials of the discrimination reaching task. Training takes place only if the arm executes an action within the "go" period: in this case if it executes the correct action, it receives a reward R=1, otherwise it receives a reward R=0. The actor-critic network is trained with an error backpropagation algorithm on the basis of the state $\mathbf{x}_t$ of the input layer at time t, an *error* for *all* the ten output units equal to S_{t+1}, and a different learning rate equal to 0.1 for the evaluator unit, 0.025 for the unit corresponding to the executed action, and 0 for all other actions. This learning process has the following desired effects: (a) the evaluator's evaluations V of the perceived states $\mathbf{x}$ progressively approaches the average reward R obtained by the actor in such a state; (b) the signal sent to the decision maker's accumulator unit corresponding to the executed action is increased if S>0 in correspondence of $\mathbf{x}$ (so that this action will have higher chances to win the race when $\mathbf{x}$ is encountered again) while it is lowered if S<0; (c) the signal sent to the other actions' accumulators is not changed.

Results

Training of the first version of the posture controller during the "childhood learning phase" was partially successful: the error decreased from about 14 cm to about 4 cm; training the second version of it led to better results: the error decreased to about 0.5 cm (Figure 2): this performance is quite good considering that only part of the 10x10 input units of the posture controller were used by this component (as the map of them had to cover the whole irregular space reachable by the arm, see Figure 2): as a result the encoding of the desired hand-position was quite sparse. Interestingly, and as hoped, in both cases the controller exhibits a *powerful generalization capability*: it is capable of reaching a target from any starting posture, even if this requires more than one step (note: this is a condition for which it has never

been trained; see Figure 2). Table 1 suggests how this is possible in the case of the posture controller's second version: the weights emerged are such that the arm segments are moved in the correct direction until the arms' angles match the desired ones. In the case of the controller's first version, the weights emerged tend to encode the desired angles in an almost direct fashion (data not shown).

Training of the system (actor-critic) for 500,000 trials in the "adulthood learning phase" was not wholly successful: in 10 different runs of the simulation with different random-number generator's seeds (these cause different initial random weights of the actor-critic network), the system learned to reach between 11 and 13 targets out of the possible 16 ones (given by eight positions x two colors; notice that the two real monkeys trained by Cisek & Kalaska, 2005, had a performance of 75% and 96%). Figure 3 shows an example of learning curve of a system that achieved a performance of about 12/16.

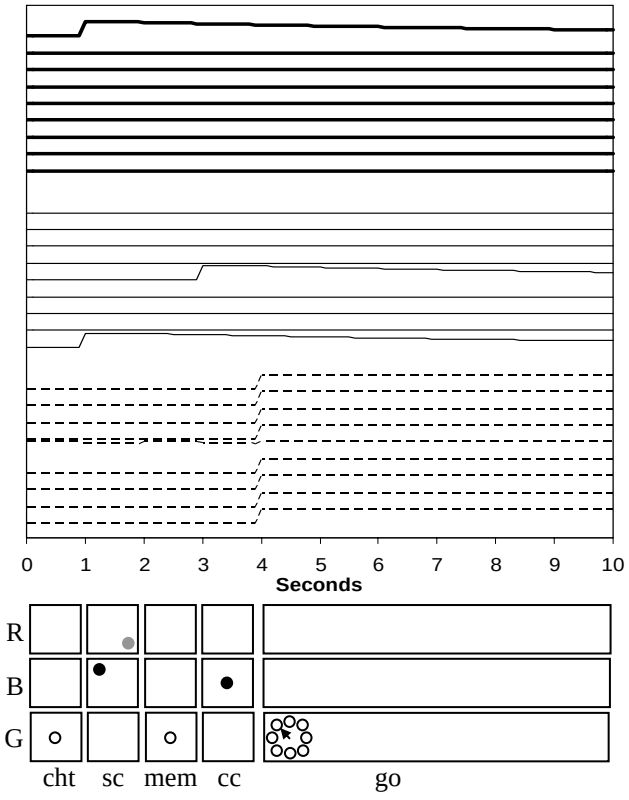

Figure 4: Top: activation during one trial of the memory units, relative to the nine positions, sensitive for green (dashed lines), blue (continuous lines) and red (bold lines). Bottom: activation of the green (G), blue (B) and red (R) screen stimuli during the trial (the boxes cover the duration time of the phases of the task illustrated in the introduction).

The analysis of the system's functioning shows that the memory units maintain a sustained activation during the task (Figure 4): this fuels the race of the accumulator units in the go phase until one of them reaches the threshold, wins the competition, and triggers the posture controller to pursue the goal-posture corresponding to it (Figure 5).

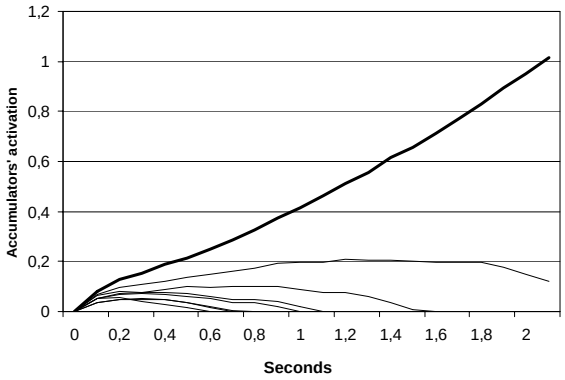

Figure 5: Activation of the accumulator units during the go phase of one trial: the goal-posture corresponding to the unit whose activation reaches the threshold of 1 is pursued by the posture controller.

Figure 6 analyses the performance of the system with respect to the nine possible target positions and the two possible targets' colors. The graph and direct analysis of data indicate that the system always successfully reaches the targets 2, 3, 4 and 9, independently of their color (targest are numbered from 1 to 9 along 3 rows from left to right and from top to bottom). In contrast, it reaches targets 1, 6, 7, and 8 only when they are of either one of the two colors (the system correctly learns to avoid selecting target 5, corresponding to the hand's initial position). Remarkably, in all trials in which the system fails to reach a target, it reaches the target located at the opposite position with respect to it: this means that the system has learned to focus on the two spatial alternative targets but fails to integrate this information with the color-cue information (this is in agreement with the neural activations observed in real monkeys, cf. Cisek & Kalaska, 2005).

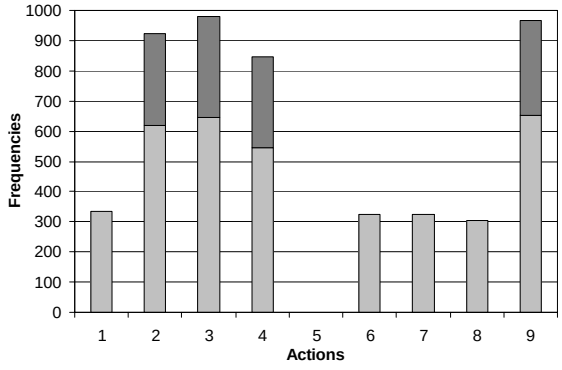

Figure 6: Number of trials out of 5000 in which the trained system selects the nine target-postures. Light-gray bars indicate successful selections while dark grey bars indicate that the selection felt on the target located at the opposite position with respect to the correct target.

These results, although preliminary in many respects (see conclusions), show that the model represents an integrated working hypothesis on the overall organization of organisms' motor system, which is computationally sound and captures several important biological mechanisms underlying it.

Conclusions and Future Work

This paper presented a model that is novel in several aspects: (a) the model presents a first integration of a (one-step) actor-critic architecture, one of the best biologically-plausible models of classical-conditioning-like tasks and basal ganglia, with an "accumulator model", one of the best biologically plausible models of decision making; (b) the model integrates this actor-critic component with a goal-based repertoire of actions that are learned by the system on the basis of self-generated experience; (c) the model presents a useful working hypothesis, in the form of a wholly integrated architecture, of the possible organization of the primitive-based hierarchical sensorimotor system of organisms.

The model has also important limitations, that will be the starting point for future work: (a) the input portion of the system is a simple 27-cell map: the system should be tested with a more realistic input component; (b) the actor-critic portion of the system is trained on the basis of the implausible "error backpropagation algorithm": can this be substituted with a more biologically plausible algorithm? (c) the experiments have shown that the model functions with 9 sensorimotor primitives: how would it scale to larger numbers? (d) the accumulator model allows selecting only discrete and locally represented actions: how is it possible to allow it to select actions represented continuously and distributely (cf. Doya, 2000)? (e) last and most important: the system is currently incapable of learning to *select actions at the right moment* after perceiving a sequence of stimuli: how is it possible to overcome this limit?

Acknowledgments

This research has been supported by the project "MindRACES - From Reactive to Anticipatory Cognitive Embodied Systems", funded by the European Commission under grant FP6-511931.

References

Arbib, M. (1981). Visuomotor coordination: from neural nets to schema theory, *Cognition and Brain Theory, 4*, 23-39.

Baldassarre, G. (2002). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. *Journal of Cognitive Systems Research, 3,* 5-13.

Baldassarre, G., & Parisi, D. (2000), Classical and instrumental conditioning: from laboratory phenomena to integrated mechanisms for adaptation. *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior - Supplementary Volume* (pp. 131-139). Honolulu, USA: International Society for Adaptive Behavior.

Barto, A.G.; Sutton, R.S., & Anderson, C.W. (1983). Neuronlike adaptive elements that that can learn difficult control problems. *IEEE Transactions on Systems Man and Cybernetics, 13,* 835-846.

Cisek, P., & Kalaska, J. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, *45*, 801-814.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation, 12*, 219-245.

Giszter, S.F., Mussa-Ivaldi, F.A., Bizzi, E. (1993). Convergent force fields organised in the frog's spinal cord. *Journal of neuroscience, 13(2),* 467-491.

Graziano, M.S., Taylor, C.S., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron, 34,* 841-851.

Houk, J.C.; Davis, J.L., & Beiser, D.G., (Ed.) (1995). *Models of Information Processing in the Basal Ganglia.* MIT Press, Cambridge, MA.

Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (2000). *Principles of Neural Science.* New York: McGraw-Hill.

Kuperstein, M. (1988). A neural model of adaptive hand-eye coordination for single postures. *Science, 239*, 1308-1311.

Rainer, W.P., & Tani, J. (2004). Motor primitive and sequence self-organisation in a hierarchical recurrent neural network. *Neural Networks, 17*, 1291-1309.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back--propagating errors. *Nature, 323*, 533-536.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences, 3,* 233-242.

Schall, J.D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience, 2,* 33-42.

Schultz, W., Dayan, P., & Montague, R.P. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Shadmehr R., & Wise, S. (2005). The computational neurobiology of reaching and pointing. Cambridge: MIT Press.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Swanson, L.W. (2005). *Brain Architecture.* Oxford: Oxford University Press.

Usher, M., & McClelland, J.L. (2001). On the time course of perceptual choice: the leaky competing accumulator model. *Psychological Review, 108,* 550-592.

von Hofsten, C. (1982). Eye-hand coordination in newborns. *Developmental Psychology, 18,* 450-461.