CORE
Metadata, citation and similar papers at core.ac.uk
Provided by PUblication MAnagement

A Modular Neural-Network Model of the Basal Ganglia's Role in Learning and Selecting Motor Behaviours

Gianluca Baldassarre (gbalda@essex.ac.uk)
Department of Computer Science, University of Essex
CO4 3SQ Colchester, UK

Abstract

This work presents a modular neural-network model (based on reinforcement-learning actor-critic methods) that tries to capture some of the most-relevant known aspects of the role that basal ganglia play in learning and selecting motor behavior related to different goals. In particular some simulations with the model show that basal ganglia selects "chunks" of behaviour whose "details" are specified by direct sensory-motor pathways, and how emergent modularity can help to deal with multiple behavioral tasks. A "top-down" approach is adopted. The starting point is the adaptive interaction of a (simulated) organism with the environment, and its capacity to learn. Then an attempt is made to implement these functions with neural architectures and mechanisms that have a neuroanatomical and neurophysiological empirical foundation.

Introduction, Methodology, Empirical Evidence Addressed

What is the role that basal ganglia play in mammals' sensory-motor behaviour? When organisms have different needs/goals, sometimes they have to associate slightly different behaviours to the same perception patterns, some other times they have to associate completely different behaviours to them. This work presents some simulations that suggest that in the former case the differences are dealt with within the same sensory-motor pathway (implemented by a neural module) while in the later cases different sensory-motor pathways are selected. In fact if the behavioral response to associate to a given perception were different with different needs/goals, using the same neural synapses/ pathways would only cause interference. In this context basal ganglia could play a role in selecting different sensory-motor pathways when necessary.

This work follows a "top-down" approach, where the starting point is organisms' behaviour and learning processes (cf. Meyer & Guillot, 1990). On this purpose it presents a simulation of an organism that has different needs (signals coming from the body and indicating a physiological unbalance, cf. Rolls, 1999) or, alternatively, different goals (desired states of body-world) associated to different positions in the environment (for example we can assume that these different positions are occupied by resources that satisfy different needs). The organism learns through classical and instrumental learning (Lieberman, 1993; in Baldassarre & Parisi, 2000, these two learning mechanisms are integrated in a comprehensive actor-critic model. Cf. Barto, 1995, and Sutton & Barto, 1998, for this model) to navigate in the environment in order to reach those positions. Given this behaviour as a starting point, the work attempts to yield it by building a neural-network controller that satisfies (some of) the constraints coming from the known empirical evidence about basal ganglia. Since the starting point of this approach is to simulate sophisticated organisms' behaviours, sometimes there is no empirical data suggesting which mechanisms underlie them. In these cases some computational solutions are adopted that do not have a known empirical correspondent (they will be appealed as "arbitrary" in the rest of the paper). These solutions should be considered as a useful theoretical exercise, eventually suggesting interesting ideas to the empirical investigation, and should not be judged too severely on the basis of the neural evidence.

The anatomical and physiological evidence specifically addressed in this work is now illustrated. Chevalier & Deniau (1990) propose that a double-inhibition mechanism is the basic process of basal ganglia's functioning. They report that in some experiments where monkeys have to carry out a delayed saccade to a remembered target, some striatal cells (usually mute) are induced to fire with local injection of glutammate. The striatal discharge inhibits (via GABAergic connections) a group of cells in the substantia nigra pars reticulata (usually tonically active) that release from (GABAergic) inhibition a subset of cells of the superior colliculus responsible for the saccade. In the case of skeletal movements the double inhibition releasing mechanism is implemented by the striatum-globus pallidus-thalamus pathway. The authors report that while in rodents this mechanism is sufficient to trigger movements, in the reported experiments the execution of a saccade requires temporal coincidence of basal ganglia disinhibition with command signals from other sources. This aspect is present in the model: basal ganglia select a particular sensory-motor pathway that then yields the detailed behavioral output.

Graybiel (1998), addressing the role that the basal ganglia's neural modules play in human slow habit learning and animal stimulus response association, draws an abstract parallel between the striatum's

anatomical organization in partly interconnected zones, called "matrisomes", and the modular architecture of the neural networks of Jacobs et al. (1991). As we shall see, the computational model presented here proposes a possible way to specify such parallel.

Houk, Adams, & Barto (1995) suggest a possible correspondence between the actor-critic models' architecture and functioning (Barto, 1995; Sutton & Barto, 1998) and the architecture of the basal ganglia. In particular they propose that the circumscribed regions called "striosomes" (differently from matrisomes, they are identifiable for their chemical make-up and output connectivity) may implement the function of the critic (predicting future rewards and yielding a step-by-step reward signal in cases of delayed rewards) and the surrounding "matrix" regions may implement the function of the actor (selecting actions or, as in the model presented here, sensory-motor pathways). As we shall see, the actor-critic model is at the base of the model presented here.

Lots of other aspects of these contributions have been incorporated in the model, and will be presented in detail in the next section. The numerous brain-imaging studies of basal ganglia's role in sequence learning are not directly addressed in this paper (see Graybiel, 1998, for some references).

Scenario and Model of Basal Ganglia

The environment used in the simulations is a square arena with sides measuring 1 unit (Figure 1). The organism cannot see the boundaries of the arena and cannot exit it. Inside the arena there are 5 circular landmarks/obstacles that the organism can see with a one-dimension horizontal retina covering 360 degrees with 50 contiguous sensor units. Each unit gets an activation of 1 if a landmark is in its scope, of 0 otherwise. The signals coming from the retina are aligned with the magnetic north through a compass. Before being sent to the controller, these signals are re-mapped into 100 binary units representing the image "contrasts". Two contiguous retinal units activate one contrast unit if they are respectively on and off, another contrast unit if they are respectively off and on, no contrast units if they are both on or both off (cf. Figure 1). At each cycle of the simulation the organism selects one of eight actions, each consisting in a 0.05 step in one of eight directions aligned with the magnetic north (north, northeast, east, etc.). The outcome of these actions is affected by a Gaussian noise (0 mean, 0.01 variance). The organism's task is to reach one of the three goal positions showed in Figure 1.

Figure 2 illustrates the main features of the organism's controller and the possible brain areas and nuclei corresponding to the model's components. Now a computational description of the controller is given, and its possible links to the mammal brain's neural structures are illustrated (notice that the units used in the model sometimes represent whole neural assemblies and at other times single units).

The matcher is a (arbitrary) hand-designed network responsible for generating an internal reward signal r by detecting the similarity between the goal and the current input contrasts (a goal is the contrasts' pattern at that goal position). When these patterns have at least 94% of bits with same value, the matcher returns 1 otherwise it returns 0. It is assumed that some memory process, not simulated in the model, evokes the goal patterns (when a goal is reached, another goal is evoked that is randomly chosen between the three goals). In real brains, goal patterns may be generated within frontal areas (e.g. by the frontal eye fields in the case of saccades) and recognition could take place here or in the sensory areas themselves (cf. Kosslyn, 1999).

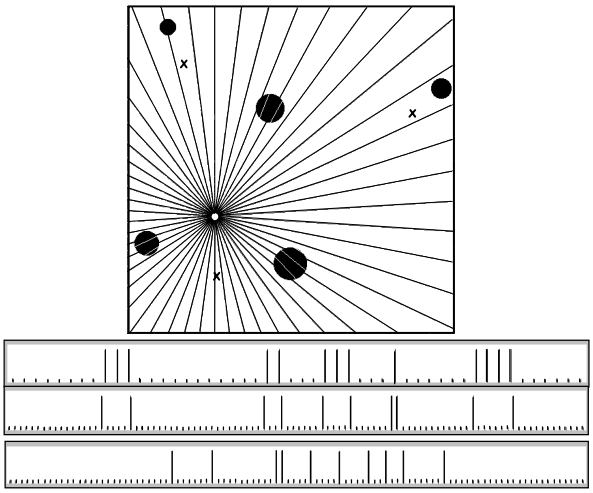

Figure 1: (Top) the scenario of the simulations containing three goals (marked with x), five landmarks (black circles), the scope of the organism's 50 visual sensors (delimited by the rays), and the organism (white circle at origin of rays). (Bottom) the activation of the visual sensors, its re-mapping into contrasts, and the bottom left goal (contrast pattern).

There is an alternative way to view this part of the model. Animals are endowed with innate neural structures that take input from the environment and map it into a "reward" or "punishment" internal signal. This usually happens when some states of the environment are achieved that are relevant for adaptation, for example some food is ingested or the body is hurt (primary reinforcements). Notice that these signals are produced only if a correspondent appetitive need (e.g. hunger) is present (Rolls, 1999). In the model the presence of a certain need could be thought of as corresponding to an arbitrary pattern (the "goal") coming from the body, while the signal relevant for adaptation is the signal coming from the sensors (e.g.

from the sensors in the mouth that detect the ingestion of food). In this case the matcher would yield a rewarding signal when a need and the corresponding satisfying input pattern are present together (in this case the matcher would correspond to limbic structures, cf. Rolls, 1999). In both cases the matcher's signal arrives to the substantia nigra pars compacta and ventral tegmental area, capable of generating a dopaminergic signal that triggers learning.

The actor, with the 6 "expert" networks (6 different input areas - thalamus - frontal areas pathways), implements the organism's "action-selection policy". Each expert is a two-layer feed-forward neural network that gets the goal and the visual contrasts as input, and has 8 sigmoidal output units that locally encode the actions (the experts may correspond to thalamus' neural assemblies: here the details of the model are quite arbitrary). To select one action, the activation m_k (interpretable as "action merit") of the output units is sent to the frontal areas where a stochastic winner-take-all competition takes place (cf. Hanes & Schall, 1996, on this possibility). The execution of one action has to be thought of involving the activation of a particular muscle template. The probability P[.] that a given action a_k becomes the winning action a_w (to execute) is given by: $P[a_k = a_w] = m_k / \Sigma_f m_f$. The role of the basal ganglia is to select an expert which, in its turn, has to select the actions to be executed through the mechanism of double inhibition illustrated previously involving the matrix of the striatum and the globus pallidus. This is done with another winner-take-all competition analogous to the previous one, but this time involving the experts instead of the actions (could this mechanism correspond to the bistable behaviour of the striatum spiny cells?). Notice that the basal ganglia can only release the proper expert from inhibition, but cannot trigger an action directly.

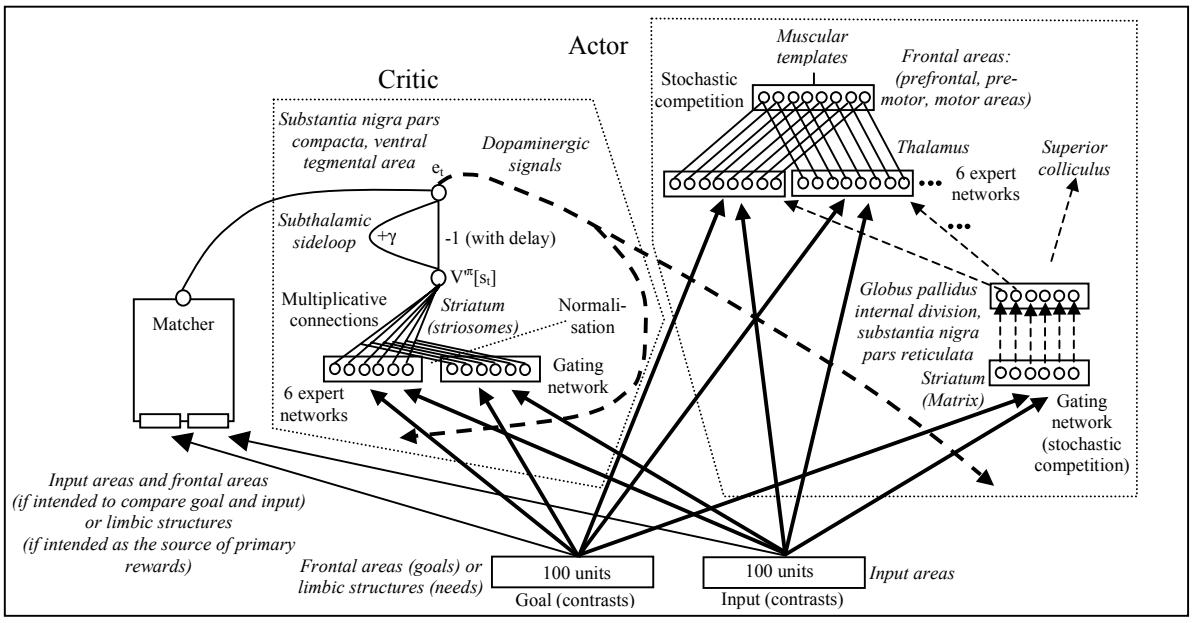

Figure 2: The components of the organism's controller. Labels in *Italic* indicate the possible brain areas and nuclei corresponding to the model's components. Thin arcs indicate one-to-one connections with weight +1 when not differently indicated. Dashed thin arrows indicated unit-to-unit/area inhibitory connections (strong enough to make the target units/areas silent). Bold arrows indicate connections updated on the basis of the dopaminergic signal. Dashed bold arrows indicate the dopaminergic signal.

The critic is a "mixture of experts network" (Jacobs et al., 1991) based on 6 expert networks. Each expert is a two-layer feed-forward neural network that gets the goal and the visual contrasts as input and has one linear output unit. The critic learns to yield the estimation $V'^{\pi}[s_t]$ of the "evaluation" $V^{\pi}[s_t]$ of the current contrast pattern s_t. $V^{\pi}[s_t]$ is defined as the expected discounted sum of all future reinforcements r, given the current action-selection policy π expressed by the actor: $V^{\pi}[s_t] = E[\gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+3} + \ldots]$ where $\gamma \in (0, 1)$ is the discount factor, set to 0.95 in the simulations, and E[.] is the mean operator. In order to compute $V'^{\pi}[s_t]$ the output v_k of the experts is weighted and summed: $V'^{\pi}[s_t] = \Sigma_k[v_k\ g_k]$, and the weight g_k is computed as the softmax activation function of the output units o_k of the gating network: $g_k = \exp[o_k]/\Sigma_f \exp[o_f]$. This part of the model is arbitrary, but the modularity of the striosomes confers some plausibility to the model (the model is an

implementation of what is suggested in Houk et al., 1995: different striosomes may be specialized in dealing with different behavioral tasks. As we shall see, this is obtained as an emergent feature of the model). The last component of the critic (subthalamic loop, substantia nigra pars compacta and ventral tegmental area) is a neural implementation of the computation of the "temporal-difference error" e defined as: $e_t = (r_{t+1} + \gamma V'^{\pi}[s_{t+1}]) - V'^{\pi}[s_t]$ (Houk et al., 1995). Each critic's expert has a specific error defined as: $e_{kt} = (r_{t+1} + \gamma V'^{\pi}[s_{t+1}]) - v_k[s_t]$. These error signals correspond to the dopaminergic signals and are at the base of the learning processes of the actor and critic.

Each critic's expert is trained on the basis of the expert's dopaminergic error signal that assumes the role of error in the estimation of $V'^{\pi}[s_t]$ in a supervised learning algorithm. The weights of the experts are updated so that their estimation $v_k[s_t]$ tends to be closer to the target value $(r_{t+1}+ \gamma V'^{\pi}[s_{t+1}])$. This target is a more precise evaluation of s_t because it is expressed at time t+1 on the basis of the observed r_{t+1} and the new estimation $V'^{\pi}[s_{t+1}]$. The formula (a modified Widrow-Hoff rule, cf. Widrow & Hoff, 1960) to update the weights of each expert is: $\Delta w_{ki} = \eta\, e_{kt}\, y_i\, h_k$ where w_{ki} is a weight of the expert, η is a learning rate (set to 0.01 in the simulations) and y_i is the activation of the goal and contrast units. h_k (absent in the Widrow-Hoff rule) is the (updated) contribution of the expert k to the global answer $V'^{\pi}[s_t]$, and is defined as: $h_k = g_k\, c_k / \Sigma_f\, [g_f\, c_f]$, where c_k is a measure of the "correctness" of the expert k defined as: $c_k = \exp[-0.5\, e_{kt}^2]$. The gating network weights z_{ki} are updated to increase the weight in yielding $V'^{\pi}[s_t]$ of the experts who had low errors: $\Delta z_{ki} = \xi\, (h_k - g_k)\, y_i$ where ξ is a learning rate set to 0.1 in the simulations. This algorithm leads the experts to specialize in the different regions of the goal-contrast space. Notice that ξ is higher than η. This has been found to be a necessary condition for the controller to work. With $\xi = 0.01$ the experts did not specialize and interference between different goals prevented learning.

The actor is trained according to the dopaminergic signal e_t. In this case this signal is interpreted as the actor's capacity to select actions that bring the organism to new states with an evaluation higher than the average evaluation experienced previously departing from that same state. The updating of the action merits of the selected expert (and only this) is done by updating the weights of the neural unit corresponding to the selected action a_w (and only this) as follows: $\Delta w_{wi} = \zeta\, e_t\, (4\, m_w\, (1 - m_w))\, y_i$. ζ is a learning rate (0.01) and $(4\, m_w\, (1 - m_w))$ is the derivative of the sigmoid function multiplied by 4 to homogenize the size of the learning rates of the actor and the <u>linear</u> critic. The model's dopaminergic signal affecting the sensory-motor pathways may correspond to the real brain dopaminergic signal targeting the frontal areas downstream the thalamus. For simplicity in the model these dopamine-sensitive areas have been designed upstream the thalamus. The weights of the winning gating network's unit are updated in the same way used for the experts' merits (learning rate 0.01).

The learning mechanism of the critic and the actor differ because in the later case it is not possible to have a teaching pattern to implement a supervised learning algorithm (as in the former case). The stochastic nature of the actor is necessary to produce new behaviours that are then strengthened or weakened according to their outcome in terms of rewards. At the beginning of the simulations the weights of the critic and actor (only those affected by the dopamine) are randomized in the interval [-0.001, +0.001]. This implies that the evaluations expressed by the linear critic are around 0, and the merits (probabilities) expressed by the "sigmoidal" actor (stochastic selector) are around 0.5 (0.125). This implies that initially, the organism's behaviour is a random walk. Then the critic and the actor are trained simultaneously (policy iteration): the evaluator learns to evaluate the states of the world on the basis of the actor's action-selection policy, and the actor improves the policy by increasing the probabilities of those actions that yield an evaluation higher than the expected one (cf. Sutton & Barto, 1998).

Simulations, Results, Interpretations

As mentioned, the task of the organism is to reach one of the three goal positions shown in Figure 1. When a goal is reached a new one (randomly chosen between the three goals) is assigned to the organism and this has to reach it from its current position. Figure 3 shows the organism's learning curve in terms of number of steps taken to reach a goal (mobile average for 100 successes, average for 10 random seeds). The performance improves from about 1000 to about 30 steps.

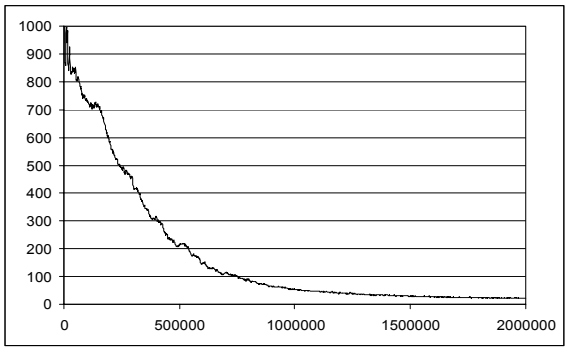

Figure 3: The learning curve of the organism. Y-axis: cycles per success. X-axis: cumulated cycles.

Figure 4 presents some data about how the neural-network controller of one of the 10 simulations has self-

organized during learning (the other random seeds have produced results with analogous quality). Concerning the critic, we see that each goal is dealt with by a different expert (in each possible position of the arena the weight of this expert in determining the evaluation is over 0.99. The second column of Figure 4 shows the resulting gradient field of the evaluations for the three goals). This probably means that the positions in the arena need to receive a different evaluation for the three different goals, so that using the same weights (same expert) would only cause negative interference. This also means that the connections from the (contrast) input pattern to the critic's gating network are redundant. The fact that different parts of the striosomes specialize for different goals as in the model, is an interesting hypothesis that has not yet been verified empirically. Notice that the controller is capable of not using some of the resources available (expert 1, 3, 4). These resources could be used for other goals.

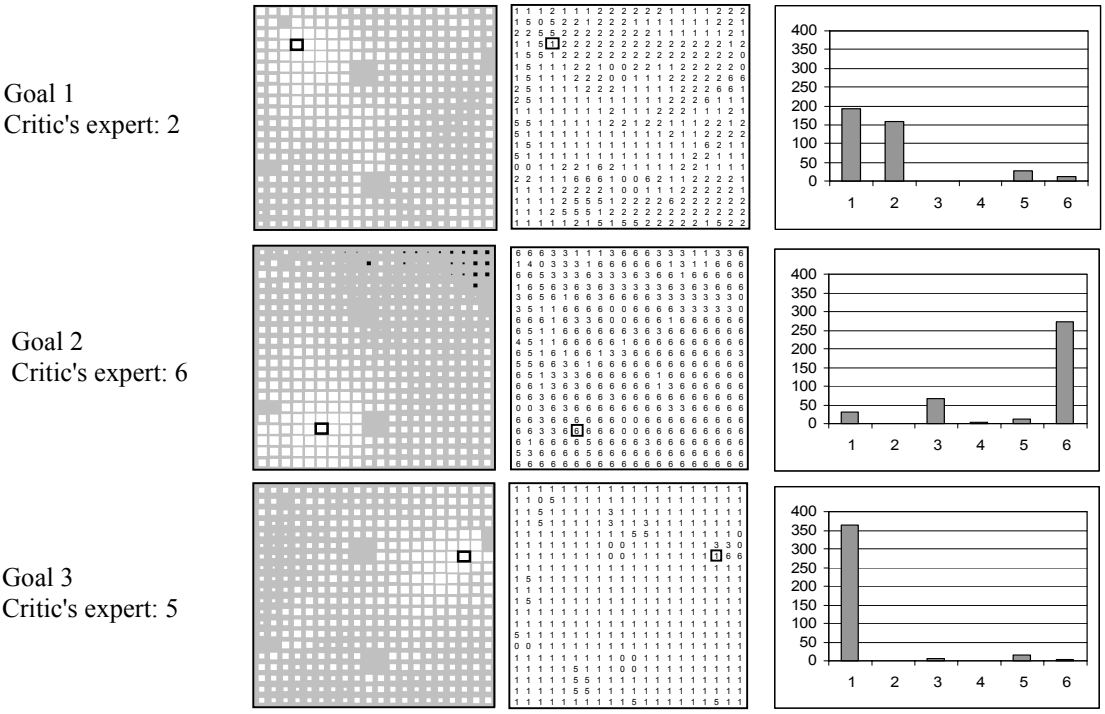

Figure 4: Data about the self-organization of the controller during learning (1 out of 10 random seeds). The three rows of graphs are relative to the three different goals. The first column reports the expert that is used by the critic for the particular goal (only 1 expert per goal). The second column of graphs reports the gradient field of evaluations $V'^{\pi}[s_t]$ yielded by the critic in 400 different positions (corresponding to the 20×20 cells of the grid). The area of the white (positive evaluations) and black (negative evaluations) cells is proportional to the evaluation yielded. The third column of graphs reports the order number of the actor's expert with highest probability of being selected (for the same 400 positions of the previous column). The last column of graphs reports the histograms that summarize the frequencies of the experts illustrated in the previous column.

With regard to the actor, Figure 4 shows that the specialization of the experts is much less pronounced. In particular the graphs of the third and fourth column of the Figure 4 show that while pursuing a goal the actor uses different experts in different position in the arena. The histograms report the frequency of use of the different experts for the different goals. Clearly the controller tends to use different experts when dealing with different goals, but now (differently from what is observed in the critic) the visual input plays an important role. An interesting fact coming out from the third and fourth column of Figure 4 is that the same experts are being used for different goals (e.g. expert 1 for goal 1 and 3). Further investigation should show if this different use of the experts in the critic and in the actor are due to the differences in the role they play or if it is due to the difference between the algorithms employed (supervised learning and stochastic unsupervised learning; cf. Calabretta et al., 1998, on the evolutionary emergence of modular networks' function through genetic algorithms). Notice that in the actor, as in the case of the critic, there is a partial use of the resources available (marginal role of expert 3, 4, and 5).

The exploration of some parameters and simulation conditions has shown some limits of the controller. Too high learning rates (especially for the critic) produce

instability, while too low rates produce slow learning. The system is also quite sensitive to the "aliasing" problem (this is the problem that occurs when there are states of the world that appear to be the same or very similar, cf. Whitehead & Ballard, 1991). In particular if there are positions that are similar to the goal positions, the organism tends to waste time searching on them (this happens because they will tend to have a high evaluation). With more goals some problems also occur: with some random seeds the same critic's expert is used for more than one goal. This produces a gradient field with more than one peak. This causes the organism to pursue the positions corresponding to these peaks at the same time so that the behaviour results to be dithering.

Conclusion

This work has presented a computational model that attempts to summarize in a coherent picture some of the most relevant properties of basal ganglia regarding motor behaviour. An attempt has been made to design a model that on one side is capable of controlling an organism in a non-trivial behavioral task, and on the other side is based on architectures and mechanisms possibly grounded on the empirical evidence about the anatomy and physiology of basal ganglia. The model has shown that the role of the striosomes in the striatum might be that of producing an evaluation of the expected future rewards, and to build a dopaminergic signal corresponding to previously neutral input patterns on the basis of some primary reinforcers. The dopaminergic signal is used to learn to express the evaluations themselves on the basis of a supervised learning algorithm. The simulations have shown that the modularity of the striosomes is used to deal with different behavioral tasks the organism meets during its life. The model has also shown that the role of the matrix in the striatum might be that of learning to generate stochastic variants of behaviour, eventually consolidated on the basis of the dopaminergic signal. Here the role of the basal ganglia's double-inhibition mechanism is not that of directly triggering particular patterns of behaviour, but that of releasing from inhibition sensory-motor pathways that then yield a particular behaviour suitably related to the current goals and percepts.

Acknowledgments

The Department of Computer Science, University of Essex, funded the author's research. Special thanks are expressed to Prof. Jim Doran (University of Essex) and Domenico Parisi (Italian National Research Council) for their valuable contribution of ideas, and Anthony Pounds-Cornish for his precious help in the preparation of the article.

References

Baldassarre, G., & Parisi, D. (2000). Classical Conditioning in Adaptive Organisms. *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behaviour - Supplement Volume* (pp. 131-139). Honolulu: International Society for Adaptive Behaviour.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In C. J. Houk, L. J. Davis & G. D. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*. Cambridge, Mass.: The MIT Press.

Calabretta, R., Nolfi, S., Parisi, D., & Wagner, G. P. (1998). Emergence of functional modularity in robots. *From Animals to Animats 5: Proceedings of the 5th International Conference on the Simulation of Adaptive Behaviour* (pp. 497-504). Cambridge, Mass.: The MIT Press.

Chevalier, G., & Deniau, M. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, 13, 277-280.

Graybiel, A. M. (1998). The basal ganglia chunking of action repertoires. *Neurobiology of Learning and Memory*, 70, 119-136.

Hanes, D. P., Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, 274, 227-230.

Houk, C. J., Adams, L. J., & Barto, G. A. (1995), A model of how the basal ganglia generate and use neural signals that predict reinforcement. In C. J. Houk, L. J. Davis & G. D. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*. Cambridge, Mass.: The MIT Press.

Jacobs, R.A., Jordan, M. I., Nowlan, S.J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79-87.

Kosslyn, S. M. (1999). *Image and Brain*. Cambridge, Mass.: The MIT Press.

Lieberman, A. D. (1993). *Learning - Behaviour and Cognition*. Pacific Grove, Ca.: Brooks/Cole Publishing.

Meyer, J.-A., & Guillot, A. (1990). Simulation of Adaptive Behavior in Animats: Review and Prospect. *Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 2-14). Cambridge, Mass.: The MIT Press.

Rolls, E. (1999). *Brain and Emotion*. Oxford: Oxford University Press.

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

Whitehead, S. D., & Ballard D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, 7, 45-83.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record, Part IV* (pp. 96-104).