A neural model of frontostriatal interactions for behavioural planning and action chunking

Nicola De Pisapia[*], Nigel H. Goddard

Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh,
5 Forrest Hill, Edinburgh EH1 2QL, United Kingdom

Abstract

Neurobiology and neuropsychology interpret the Prefrontal Cortex as a key region, in connection with others, for behavioural planning, i.e. for the generation and evaluation of goal oriented complex courses of actions, without actually executing them. We illustrate an abstract reward-based computational model of planning and its possible neural counterparts. We then focus on how the Dorsolateral PFC and the Striatum learn incrementally to chunk sequences of actions, thus allowing fast and hierarchically structured planning. We test the abstract model by simulating internal model formation and action chunking in motor and cognitive planning problems.

Keywords: Prefrontal cortex, Striatum, Dopamine, Planning, Chunking

The Prefrontal Cortex (PFC) and the Basal Ganglia (BG) are known to play a fundamental role in cognitive control and in the organization of complex behaviour [6][2]. Within this view, we emphasize that the PFC is crucially involved in generating and selecting purposeful complex sequences of actions, i.e. in behavioural planning of how to reach goals in a changing environment. In this paper we propose an abstract computational model of how the PFC, together with the BG, might achieve a planning capacity, and show two simulations.

1. Temporal Difference learning and Temporal Abstraction

The BG, and in particular the Striatum, which constitutes the input nucleus receiving connections from the PFC, are known to be involved in the acquisition of sensory motor habits and behavioural repertoires [2]. Experimental data show a high degree of modifiability in the corticostriatal connections, and the key role of the dopaminergic signal coming from the Substantia Nigra Pars Compacta (SNPC) into the Striatum, which acts as an error in the prediction of rewards. This signal has led researchers to construct computational models of the BG using Reinforcement Learning (RL) [13], and in particular actor/critic models using Temporal Difference (TD) learning, e.g., [9]. TD models in RL learn to associate states of the environment with a value function: $V^{\pi}(t)=E[r(t)+\gamma r(t+1)+\gamma^2 r(t+2)+\ldots]$, which gives an expectation at time t of the present and future rewards r, according to a discount factor $\gamma<1$, and according to a mapping of states into actions π, called a *policy*. The value function allows the system to select which actions lead to the maximal reward, using an error signal in the expectations of the rewards: $\delta(t) = r(t) + \gamma V^{\pi}(t+1) - V^{\pi}(t)$, which corresponds to the signal associated

[*] Corresponding author. *E-mail address*: pisapia@anc.ed.ac.uk

with the dopaminergic activity from the (SNPC), and which also allows the construction of sequences of actions.

To investigate planning in mammals, we make the further central hypothesis that a motor sequence, if repeatedly successful in achieving its goal, is chunked into a higher order action. Chunking has been studied in Artificial Intelligence in a variety of guises (e.g., SOAR [4]). Here we adopt TD learning together with Temporal Abstraction (TA), a recent theoretical development in RL (see [8] for more details).

TA is a mathematical description of how to construct hierarchies of chunks, called *options*. Options are defined by the set of states *I* in which the option applies, a decision rule μ which specifies what actions are executed by the option, and a completion function β which specifies the probability of completing the option on every time step. Note that also single actions can be treated as options, and that two options *a* and *b* can also be composed to produce a new option, *ab*.

To allow planning, RL agents must predict what happens when one or another option is executed over temporally abstract courses of actions. Expectations of reward and state at the end of the execution of an option *o* are given by:

$$r_s^o = E[r(t+1) + \gamma r(t+2) + ... + \gamma^{(k-1)} r(t+k) \mid \varepsilon(o,s,t)]$$

$$p_{ss'}^o = \sum_{k=1}^{\infty} p(s',k)\gamma^k, \forall s' \in S,$$

where *t+k* is the random time of termination *o, ε(o, s, t)* is the event of *o* being started in state *s* at time *t*, *S* is the set of all the states, and *p(s',k)* is the probability that *o* terminates in *s'* after *k* time steps. Note that $p_{ss'}^o$ is a combination of the likelihood that *s'* is the terminal state for *o* (e.g., information about *what* events to expect and *where* in the environment), together with a measure of how delayed that outcome will be, relative to γ (*when* to expect the end of an option to happen). These two predictions, one for the reward and the other for the final state, are called a *multi-time model* of an option.

The new extended definition of the value function becomes:

$$V^{\mu}(t) = E[r(t+1) + ... + \gamma^{k-1} r(t+k) + \gamma^k V^{\mu}(t+k) \mid \varepsilon(\mu,s,t)]$$

where *k* is the duration of the first option selected by μ. Methods used to learn value functions and selection of the optimal actions, such as TD learning, can now be generalized to learn value functions over models of options.

Given these definitions, one simple method for an agent to plan, i.e. to simulate and select options to be undertaken to achieve goals, is *Synchronous Value Iteration* [13] [8]. This consists of a step by step evaluation process of the states of the environment, that starts from the goal state, until it reaches the current state of the agent. After this process, the best current policy can be easily chosen by the agent by taking options that in each time step lead to the higher evaluation.

The key insight here is that by allowing TA, a RL agent can evaluate not only what is the state and the reward after a primitive action is taken, but also which is the state and the reward after a sequence of actions, acting at varying time scales, is taken.

We connect neural responses to our abstract account of planning by modeling modules that build options, learn their internal models and plan as just described.

2. The general model of planning

The modules of our model are motivated by the RL computational requirements for planning, as described in the previous section. Our neuroanatomical hypotheses map each of these functional modules to a specific brain area. While these mappings are supported by experimental evidence, we are well aware of the possibility that any or all of these functions may be subserved by a *network* of brain regions. Nevertheless, we hypothesise:
(1) An Internal Model module, corresponding to Dorsolateral PFC, which updates (multi-time) internal models of the world and maintains them in working memory. These models give information, precisely as in the RL description seen in section 3, about *where*, *what* and *when* to expect goal related events to happen. The *where* and *what* aspects have been largely studied in monkeys (see [1] [9] for example). The *when* aspect is consistent with findings by Watanabe [16] (see next section for a simulation).
(2) An Anticipated Rewards module, corresponding to Orbitofrontal [15] and Anterior Cingulate Cortex [5], which actively maintains goal representations, and therefore evaluates task relevance of different chunks.
(3) An Action Chunks module, corresponding to Striatum in the BG, whicy incrementally chunks the actions, under the directives of the previous modules, and organizes these chunks following a hierarchical constructive development. The idea of chunking in the BG has been suggested in [2], and tested in interesting works on the dorsolateral Striatum of rats [3].
(4) External inputs to the model are given by high level information about space and object identity in the Perception States module, corresponding to Parietal and Temporal Cortex.
(5) A Prediction Error module which allows learning by sending an error in the prediction of reward occurrence and of task related stimuli. This mimics phasic dopaminergic effects respectively from the SNPC to the Striatum, and from the Ventral Tegmental Area to the PFC [14]. A reward signal is sent at the time of occurrence of a reward to the chunking module [12].
(6) Finally an Actions module, which stores the actual actions that define each chunk. This corresponds to Motor Cortex, and also the Cerebellum. This module consolidates internal model representations of the environment once the activation based models in the Internal Model module have been several times successful in its predictions. The shift of the neural activity from frontal areas of the brain to posterior areas has been experimentally verified in motor tasks (for example in [10]).

3. Two simulations

Here we report the results of two simulations, implemented to test parts of the illustrated abstract model. The actual implementations are simple connectionist networks as in [7]. The units are binary, and in different modules they can represent single stimuli, single actions or single options. Learning is based on TD and TA.

First we tested our model to see if the construction of multi-time internal models could match cellular recordings in Dorsolateral PFC. We simulated Watanabe's task [16] to look into temporal aspects of the construction of internal models. In his experiments, the animals (*Macaca fuscata*) had to learn a delayed response task, where they received a specific food if, after a short left or right light presentation, a fixed delay, and a go signal, they pressed a left or right button, depending on where the light was presented. The specific food kept fixed in each session of 50 trials, but then modified in subsequent sessions. Some of the neurons in Dorsolateral PFC displayed spatial specificity, namely they showed activity related to which side the light was presented. Some other neurons, as the one simulated here, didn't show this spatial specificity, but they were related to the specific kind of food that the animal was expecting. These Dorsolateral PFC neurons showed, after training, an anticipatory activity, which started at a low level at the beginning of the task, grew maximally at the expected time of specific food occurrence, and then decayed rapidly. We interpret this activity as the neural codification of multi time models, namely it is a combination of the likelihood that a specific food is the terminal state for an action, together with a measure of how delayed that outcome will be, as reported in section 1.

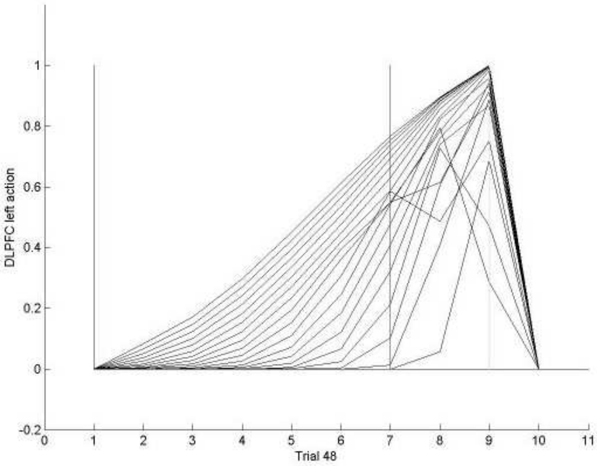

Fig. 1 The learning of the expectation of *when* specific food will be delivered in a simulation of [15].

In the simulation we divide the task into 11 discrete time steps. The three vertical lines in Fig. 1 indicate from the left respectively cue onset (light on the left, in this case) (step 1), go signal (step 7), and food delivery (step 9). In the figure we see activity of a unit, corresponding to the prediction relative to the probability of terminating *press-left* actions with the specific food. What is showed is actually the learning process of this prediction, trial after trial. The higher curves show the prediction in later trials, and therefore are those that must be compared with Watanabe's plots. As in the biological recordings, in this unit we see that the expectation of option termination grows steadily until the time of delivery.

In the second simulation we test the chunking capacity of our general model. An agent plays the Tower of Hanoi, a game used in cognitive psychology to study planning capacities in humans. There are five disks of different size and three pegs. Any disk can be moved to any empty peg or onto any other disk of bigger size. Given any starting configuration, the task is to reach a goal configuration in the shortest number of moves. In our simulation initially the agent is capable only of basic actions, and therefore of planning only about simple goals. But as learning takes place, the agent chunks sequences of actions, and therefore plans also about more and more complex (distant) goals.

In the plot in Fig. 2 we compare performance by first generating a sequence of 60 tests (in the form of start/end pairs); we then run the simulation on this specific sequence both with and without the chunking mechanism. What we report here is the number of steps it takes to the evaluative planning process, as in *Synchronous Value Iteration* (see section 1), to choose the best policy.

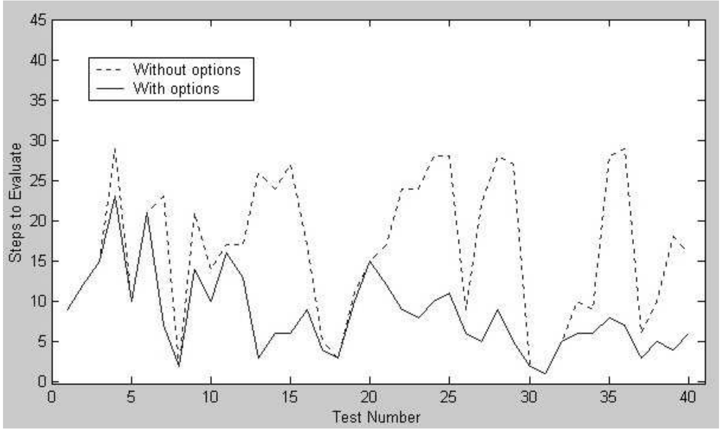

Fig. 2 The Tower of Hanoi simulation using planning and chunking.

The chunking process improves performance considerably during the test sequence, as the model learns chunks from earlier tests (lower numbers). These chunks are used in later tests to plan more efficiently.

4. Discussion

This general reward based model, composed an internal model construction module, an action chunking mechanism, and the capacity to shift action representations to other modules during skill acquisition, is the first attempt to account for actual cellular recordings in Dorsolateral PFC and also computational efficiency in planning.

Several modifications to the general model can be imported. For example, different ways to learn internal models using TA can be configured. The described methods require an option to be executed until termination, others, called *intra-option*, where learning that takes place *off-line*, learn about an option from small fragments of experience consistent with that option, even if the option itself is not executed. In these methods, options not executed and not relevant to the reward that the system is trying to maximize, can nonetheless be improved for a future use.

Another modification regards when the system should actually chunk a sequence of actions. This is currently a major area of research in RL [8]. In the simulation reported here the chunking mechanism comes into play indiscriminately, every time a sequence of actions is successful in achieving its goal. Even this indiscriminate mechanism improves computationally the performance, but more sophisticated procedures can be designed.

Current work is focussing on implementing these modifications, extending the simulations to other parts of the model, and on designing animal and human experiments to further test the hypotheses. Future work will concentrate on enriching the biophysical aspects of the model.

References

[1] P. S. Goldman-Rakic, Cellular basis of working memory. Neuron. 14 (1995) 477-485.

[2] A. M. Graybiel, The basal ganglia and chunking of action repertoires, Neurobiology of learning and memory 70 (1998), 119-136.

[3] M. S. Jog, Y. Kubota, C. I. Connolly, V. Hillegaart, A. M. Graybiel, Building neural representations of habits. Science, 286 (1999) 1745-1749.

[4] J. E. Laird, P. S. Rosenbloom, A. Newell, Chunking in Soar: the anatomy of a general learning mechanism. Machine Learning 1 (1986), 11-46.

[5] A. W. MacDonald III, J. D. Cohen, V. A. Stenger, C. S. Carter, Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science 288 (2000) 1835-1838.

[6] E. K. Miller, J. D. Cohen, An integrative theory of prefrontal cortex function, Annual Review of Neuroscience 24 (2001), 167-202.

[7] P. R. Montague, P. Dayan, T. J. Sejnowski, Framework for mesencephalic dopamine system based on predictive hebbian learning, Journal of Neuroscience 16 (1996) 1936-1947.

[8] D. Precup, Temporal abstraction in reinforcement learning, PhD thesis, University of Massachusetts Amherst, 2000.

[9] S. C. Rao, G. Rainer, E. K. Miller, Integration of what and where in the primate prefrontal cortex. Science. 276 (1997) 821-824.

[10] R. Shadmehr, H. Holcomb, Neural Correlates of motor memory consolidation. Science 277 (1997), 821-825.

[11] W. Schultz, Predicted reward signal of dopamine neurons. Journal of Neurophysiology, 80 (1998) 1-27.

[12] W. Schultz, Multiple reward signals in the brain. Nature 1 (2000) 199-207.

[13] R. S. Sutton, A.G. Barto, Reinforcement Learning: an introduction. Boston: MIT Press 1998.

[14] R. Suri, Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model, Experimental Brain Research, 140 (2001), 234-240.

[15] L. Tremblay, W. Schultz, Relative reward preference in primate orbitofrontal cortex. Nature. 398, (1999) 704-706.

[16] M. Watanabe, Reward expectancy in primate prefrontal neurons. Nature 382 (1999) 629-632.

Nicola De Pisapia is a PhD student in the Institute for Adaptive and Neural Computation, School of Informatics, at the University of Edinburgh. He received a Master in Logic and Computation from Carnegie Mellon University in 1999, and a Master in Philosophy and Artificial Intelligence from the University of Napoli in 1995. His research interests include neurocomputational models of behavioural planning.

Nigel Goddard is a Reader at the Institute for Adaptive and Neural Computation, School of Informatics, at the University of Edinburgh. He received his Ph.D. in Computer Science from the University of Rochester in 1992. His research interest include collaborative modelling and databasing software for the neurosciences, and neurally-plausible modelling of high-level cognitive function.