Using an Internal Model in a System with Dopamine-like Reinforcement Learning

Michael Arnold

Extensive work exists relating dopamine neuron activity to the reward prediction error signals in models of temporal difference learning, indicating a mechanism by which animals may learn the sequence of actions required to perform a given task. Animals also seem to make use of internal models of the world, both when learning and when planning the execution of a task. They learn to anticipate predictable events and their behaviour in the sensory preconditioning paradigm reflects the formation of novel associative chains. As in Dyna architectures, the internal model can be used as a replacement for real experience. This facilitates planning and action selection, allowing possible scenarios to be evaluated within the safety of a virtual environment. If the internal model can also generate the reward signal, then learning can proceed at the same time as this virtual search and evaluation.

It is not clear where such internal models may exist within the central nervous system nor how they could be learnt, but it would be clearly advantageous if such learning did not necessitate an explicit reward signal. Learning an internal model should be capable of occurring during the animal's normal exploration of the world, outside of the context of the execution of any specific rewarded task. Latent learning experiments consisting of several experimental phases show that animals can profit from the unreinforced exploration of a maze in an initial phase, to find the shortest route to the reward in a later phase. The cerebellum is frequently proposed as one of the sites for learning feed-forward or internal models. The sensory consequences of actions are particularly easy to formulate within this scenario because the sensory states can be directly observed indicating a solution by supervised learning. Temporal difference approaches to learning internal models have also been tried (Suri2001) where the error in reward prediction is replaced with the error in state prediction. This method is sufficient to learn novel associative chains but its biological plausability and generality are not clear.

To study the interaction between reward-based learning, planning and internal models we describe a functional system applied to the block building task, where the subject must reproduce a model within the specified work-space using blocks taken from a resource space. We first describe the system without the internal model. It is based on the use of a cortical-basal ganglia-thalamic loop for the selection and evaluation of actions. By analogy to actor-critic methods of reinforcement learning, the basal ganglia corresponds to the actor, and the ventral tegmental area corresponds to the critic. As in Dehaene2000, this can constitute an auto-evaluative loop. The basal ganglia selects an action for a given pattern of activity in prefrontal areas. As in normal reinforcement learning scenarios, initially the basal ganglia chooses blindly, but after training on the task it learns to make the correct selection. The ventral tegmental area generates reward prediction errors which are used both to drive the learning in the basal ganglia and the ventral tegmental area, and to drive the exploration in the system. Negative errors corresponding to low levels of dopamine activity, destabilise the pattern of activity in prefrontal areas and allowing a new stable pattern to emerge. This institutes a search mechanism that complements the action selection in the basal ganglia. This virtual search and evaluation proceeds until a sufficiently potentially rewarding decision has been identified by the ventral tegmental area. As learning proceeds and the basal ganglia learns the task then less virtual search by trial and evalution is required.

This basic system is extended to planning through the use of a fixed internal model to replace the external world. At this stage we do not consider how the internal model is learnt or where it is located, we only assume that it is distinct to the basal ganglia and the pre-frontal areas. At this point our concern is primarily to understand the issues involved in introducing an internal model within the system. The internal model allows a sequence of actions or a path, rather than simply a single action, to be evaluated. As in the previous auto-evaluative loop, the use of the internal model is controlled by the error signal from the ventral tegmental area. A path is explored until either it is evaluated as being sufficiently potentially rewarding, or until a negative error disrupts the pre-frontal activity causing the path to be abandoned. The introduction of an internal model raises a number of issues. Firstly, it introduces a working memory requirement to remember the candidate path as it is explored, so that it can be reproduced if selected. For simplicity we abstract this area of working memory using a stack. Secondly, it introduces the issue of how far to unwind the stack on a negative error.

A shortcoming highlighted by the current system is the inability of the internal model to guide the exploration phases. Whether evaluating a single step or using an internal model to evaluate a sequence of steps, the search mechanism is implemented wholly through the disruption of pre-frontal activity by the reward prediction error encoded by the dopamine signal. For the system described, this can be considered a blind search as opposed to a guided search. The ventral tegmental area and the basal ganglia both learn as the system is trained on the task, but the ventral tegmental area is learning something akin to a value function and is capable only of evaluating a decision. It can only guide the making of a decision through the iterative evaluation of a set of candidates. As the actor, the basal ganglia does learn to make optimal decisions, but only over many repetitions of the task. When operating in regimes where the value function is zero, the system must rely on an sequential iterative multi-step blind search until it reaches a non-zero regime, when what is required is to use the knowledge within the internal model to achieve either a parallel or a guided search.

References

R.E. Suri "Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model", Exp Brain Res 2001 140:234-240

S. Dehaene and J-P. Changeux "Reward-dependent learning in neuronal networks for planning and decision making", Prog Brain Res 2000 126:217-229