7

Reinforcement Learning Embedded in Brains and Robots

Cornelius Weber[1], Mark Elshaw[2], Stefan Wermter[3], Jochen Triesch[1] and Christopher Willmot[3]
[1] Frankfurt Institute for Advanced Studies, [2]University of Sheffield, [3]University of Sunderland
[1] Germany, [2,3] UK

1. Introduction

In many ways and in various tasks, computers are able to outperform humans. They can store and retrieve much larger amounts of data or even beat humans at chess. However, when looking at robots they are still far behind even a small child in terms of their performance capabilities. Even a sophisticated robot, such as ASIMO, is limited to mostly pre-programmed behaviours (Weigmann, 2006). The reliance on robots that must be carefully programmed and calibrated before use and thereafter whenever the task changes, is quite unacceptable for robots that have to coexist and cooperate with humans, especially those who are not necessarily knowledgeable about robotics. Therefore there is an increasing need to go beyond robots that are pre-programmed explicitly towards those that learn and are adaptive (Wermter, Weber & Elshaw, 2004; Wermter, Weber, Elshaw, Panchev et al., 2004). Natural, dynamic environments require robots to adapt their behaviour and learn using approaches typically used by animals or humans.

Hence there is a necessity to develop novel methods to provide such robots with the learning ability to deal with human competence. Robots shall learn useful tasks, i.e. tasks in which a goal is reached, if executed successfully. Reinforcement learning (RL) is a powerful method to develop goal-directed action strategies (Sutton & Barto, 1998). In RL, the agent explores a 'state space' which describes his situation within the environment, by taking randomized actions that take him from one state to another. Crucially, a reward is received only at the final goal state, in case of successful completion. Over many trials, the agent learns the value of all states (in terms of reward proximity), and how to get to higher-valued states to reach the goal.

In Section 2 we will review RL in the brain, focusing on the basal ganglia, a group of nuclei in the forebrain implicated in RL.

Section 3 presents algorithms for RL and describes their possible relation to the basal ganglia. In its canonical formulation, RL maps discretely defined states to discrete actions. Its application to robotics is challenging, because sensors, such as a camera, deliver high-dimensional input that does not define a state in a way suitable for most tasks. Furthermore, several actions are to be learnt in different contexts with different reward types being given.

Source: Reinforcement Learning: Theory and Applications, Book edited by Cornelius Weber, Mark Elshaw and Norbert Michael Mayer
ISBN 978-3-902613-14-1, pp.424, January 2008, I-Tech Education and Publishing, Vienna, Austria

In Section 4 we will address how a neural network performing RL can be embedded in a larger architecture in which other modules follow different processing and learning principles. Taking inspiration from the brain, the sensory cortex may extract meaning from sensory information that may be suitable for defining a state as it is used for RL by the basal ganglia (Weber, Muse, Elshaw & Wermter, 2005). The motor cortex on the other hand may store movement primitives that may lead from one state to the next. Moreover, the basal ganglia might delegate learnt movement primitives to the motor cortex, so to focus on the learning of other, in particular higher-level, actions (Weber, Wermter & Elshaw, 2006).
Section 5 addresses vision, an untypical field for RL. We posit that visual stimuli can act as reinforcers for saccade learning (Weber & Triesch, 2006) and gaze following, leading to the emergence of mirror neuron like representations in motor cortex (Triesch, Jasso & Deák, 2007), and altering neuron properties in visual cortical areas (Roelfsema & Ooyen, 2005; Franz & Triesch, 2007). Together, this encourages a view in which RL acts at the core, while unsupervised learning establishes the interface to a complex world.
Section 6 discusses whether experiments are based on oversimplifying assumptions.

2. Anatomy and physiology

Our focus will be reinforcement learning in the basal ganglia. However, since the basal ganglia's main outputs are inhibitory, and since they are not yet connected in neonates (Humphries, Gurney& Prescott, 2005), there must be more fundamental brain substrates for behaviour/action initiation.

2.1 Reticular formation

The brain's reticular formation (RF) has been proposed as such a device for action selection (Humphries et al., 2005; Kilmer, 1997). The RF's giant neurons receive input from many brainstem nuclei, enabling them to sample from all sensory systems, and their axons bifurcate to project downward to the spinal cord as well as upward to the midbrain, enabling the production of motor behaviour and the control of higher-level brain centers.
The RF contains several specialized circuits. A potent example are the giant neurons in the caudal pontine RF which respond at very short latency to acoustic stimuli, and which are hypothesized to elicit the startle response to a loud and unexpected acoustic stimulus (Lingenhöhl & Friauf, 2004). The paramedian pontine RF is involved in the control of horizontal eye movements, and the midbrain RF in vertical eye movements (Sparks, 2002; Weber & Triesch, 2006).

Model of Behaviour Generation

Kilmer (1997) proposed a "command computer" model of the RF which outputs one behaviour, given as input several vectors of recommended behaviours, originating from several sensory systems. The RF model computes the winning behaviour using a relatively small number of connections and a distributed representation. Humphries et al. (2005) optimized the originally randomized connectivity by a genetic algorithm. In a robotic demonstration involving the behaviours 'wander', 'avoid obstacle' and 'recharge energy', the genetic algorithm augmented the model's behaviour selection from near-chance levels to achieving very long survival times.

The RF is rarely implicated in learning (see Bloch and Laroche (1985) for a counter-example), but rather seems "pre-programmed" at birth. Other brain structures are needed to allow adaptation to beneficial circumstances in the environment.

2.2 Basal ganglia

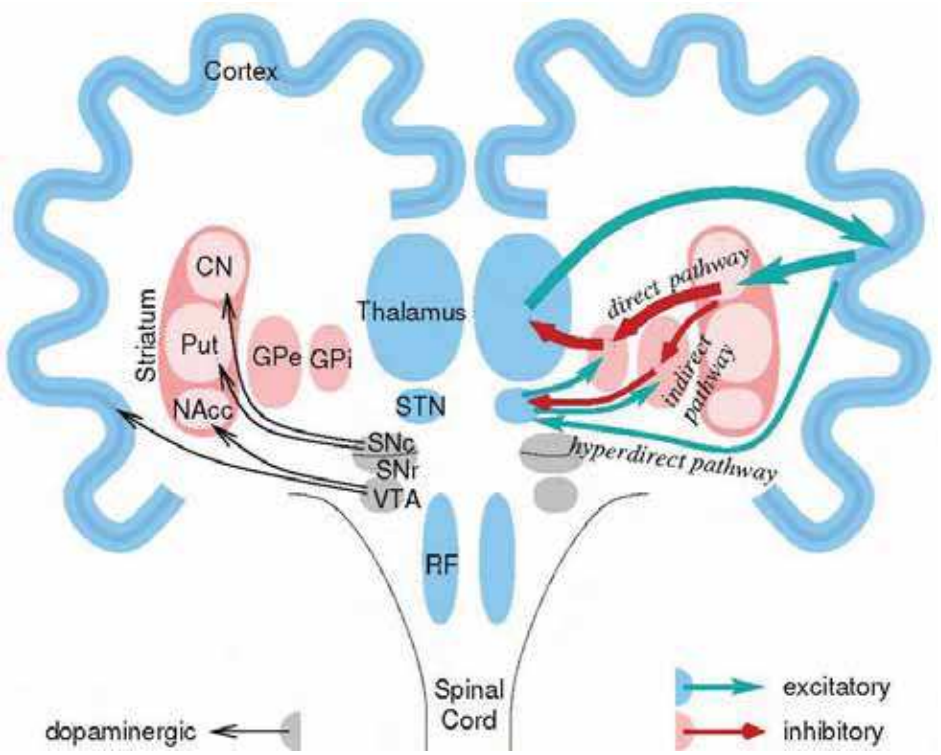


Fig. 1. Selected brain areas and connections. The thick arrows denote the primary basal ganglia (BG) → thalamus → cortex loop. This includes the direct pathway through the BG via striatum and GPi. The indirect and hyperdirect pathways are via STN, GPe and GPi. The SNr has a similar connectivity as the GPi (not shown for simplicity), so one often refers to "GPi/SNr". Dopaminergic nigro-striatal projections from SNc reach the CN and Put which make the dorsal striatum. Meso-limbic projections from VTA reach the NAcc which is part of the ventral striatum. Meso-cortical projections are from VTA to regions in the prefrontal and cingular cortex. Abbreviations: Inside the BG: CN = caudate nucleus; Put = putamen; NAcc = nucleus accumbens; GPe/i Globus pallidus externus/internus; STN = subthalamic nucleus; SNc/r = Substantia nigra pars compacta/reticulata. Outside of the BG: VTA = ventral tegmental area; RF =reticular formation.

Figure 1 shows the relevant areas and their abbreviations related to the basal ganglia (BG). The main input nucleus of the BG is the striatum which receives its main input from motor and prefrontal cortex, but also from intralaminar thalamic nuclei (not shown in Figure). The

striatum accounts for approximately 95% of the total neuron population of the BG in the rat (Wood, Humphries & Gurney, 2006). The dorsal striatum (neostriatum) consists of the putamen and the caudate nucleus. The ventral striatum consists of the nucleus accumbens (core and shell) and the olfactory tubercle (not shown in Figure). The principal neurons of the dorsal striatum, the medium spiny neurons, are inhibitory GABAergic projection neurons. They emit collaterals to neighbouring spiny neurons before they project to output stages of the BG, namely to either GPi or SNr (Houk et al., 2007).

According to Shepherd (2004), the cortical and thalamic afferents to the BG have a cruciform axodendritic pattern. This implies that individual axons cross the dendritic fields of many neurons in the neostriatum, but make few synapses with any particular cell (Wilson, 2004). The opposite is also true. Any particular neostriatal neuron can synapse (sparsely) with a large number of afferents.

Optimal Decision Making

Based on the connectivity of the BG, Bogacz and Gurney (2007) propose a model of optimal decision making that implements the statistical *multihypothesis sequential probability ratio test* (MSPRT). The underlying assumption is that the different regions of the cortex each send evidence y_i for a particular decision i to the striatum. A problem of passing this directly to the thalamus is that the action would be performed as soon as the accumulated evidence in a given channel reaches a certain threshold. This is not optimal, because in the presence of noise, a wrong channel could first reach threshold (not to mention the "technical" problem of defining when to start to integrate, as addressed in Stafford and Gurney (2007)). Rather should the difference between the favored channel and the other channels reach a threshold. Mathematically, $y_i - \ln \sum_j e^{y_j}$ is better sent to the thalamus.

Bogacz and Gurney (2007) identify the first term y_i with the *direct pathway*: the striatum inhibits the GPi/SNr which then disinhibits a corresponding thalamic region so to perform the action. The positive sign is because the tonically spiking inhibitory GPi/SNr neurons are silenced. The second term $-\ln \sum_j e^{y_j}$ represents the *indirect (hyperdirect) pathway*. It has a negative sign because the cortical afferents excite the STN (the only excitatory nucleus of the BG) which then excite the GPi/SNr neurons' inhibitory activity. Diffuse STN → GPi/SNr connections implement the sum over all channels.

This model is minimal in terms of its mechanisms, and encourages additional functionality to be implemented in the same structures. For example, the number of hypotheses y_i in the input is the same as the number of outputs; however, the BG has a much larger input structure (striatum) than output structure (GPi/SNr), which suggests a transformation to take place, such as from a sensory to a motor representation. For example, the GPi/SNr might extract a low-dimensional subspace from the high-dimensional cortical and striatal representations by a principle component analysis (PCA) algorithm (Bar-Gad, Havazelet-Heimer, Goldberg, Ruppin & Bergman, 2000). Learning, not only action selection, becomes important.

Rewards

Dopamine neuron activity in SNc and VTA is known to be correlated with rewards, learning and also with addictive behaviour. Dopamine neurons are active during delivery of an unexpected reward. If a stimulus predicts a future reward, then they will instead become active at the onset of the reward-predicting stimulus (Fiorillo, Tobler & Schultz, 2003). Dopamine neuron firing in the VTA is suppressed by aversive stimuli (Ungless, Magill & Bolam, 2004). These neurons (SNc borders the SNr, an output nucleus of the BG) project

dopaminergic axon fibres into the input nuclei (Fischer, 2003): the *nigro-striatal* projection is from the substantia nigra to the dorsal striatum; the *meso-limbic* projection is from the VTA to the ventral striatum; there is also a *meso-cortical* projection from the VTA to prefrontal and cingular cortex (Fig. 1). Consequentially, during the delay period of delayed-response tasks neurons in striatum were found to be selective for the values of individual actions (Samejima, Ueda, Doya & Kimura, 2005), and in orbitofrontal (a part of prefrontal) cortex neural activity represents the value of expected reward (Roesch & Olson, 2004). There may be a finer grain resolution of the reward delivery system, as Wilson (2004) suggests that any local region of the neostriatum may receive its dopaminergic innervation from a relatively small number of dopaminergic neurons.

The concept of a reward is however wider. Unexpected, biologically salient stimuli elicit a short-latency, phasic response in dopaminergic neurons (Dommett, Coizet, Blaha, Martindale & Lefebvre, 2005). If not reinforced, responses to novel stimuli become habituated rapidly, and the responses to rewarding stimuli also decline if stimuli can be predicted.

D1 and D2 Receptors

Dopamine has varying effects on neurons, because different neurons have different dopamine receptors. Lewis and O'Donnell (2000) state "D1 receptors may enhance striatal neuron response to [excitatory] NMDA receptor activation, where as D2 receptors may decrease responses to non-NMDA [e.g. inhibitory] receptors". Vaguely interpreted, D1 supports direct excitatory responses and D2 supports later decisions by limiting inhibition. This correlates with the findings of Hikosaka (2007) who make use of the fact that saccades to highly rewarded positions are initiated earlier than saccades to less rewarded positions. Injections of dopamine D1 receptor antagonist delayed the early, highly rewarded saccades. Injections of D2 antagonist delayed even more the later, less rewarded saccades.

The models of Brown, Bullock and Grossberg (2004) and Hazy, Frank and O'Reilly (2007) (Section 4) feature 'Go' cells which have the D1 receptor and project along the *direct pathway* to facilitate an action, and 'NoGo'/'Stop' cells which have the D2 receptor and which project to the *indirect pathway* to suppress an action.

In addition to facilitating activation, dopamine directly facilitates learning by increasing the number of synaptic receptors (Sun, Zhao & Wolf, 2005). As an example of dopamine-modulated Hebbian learning, Reynolds, Hyland and Wickens (2001) showed that synapses between the cortex and the striatum could be potentiated only with concurrent stimulation of the substantia nigra.

BG-Thalamo-Cortical Loops

The function of the basal ganglia as a learning action selection device makes sense only in the context of its main input, the cortex and its main output, the thalamus. Wilson suggests that the striatum contains a functional re-mapping of the cortex. For example, motor and somatosensory cortical representations of a single body part specifically converge on a particular region of the putamen (Flaherty & Graybiel, 1991), which is implicated in sensory guided movements.The other part of the neostriatum, the caudate nucleus, receives input from more anterior cortical areas and is implicated in memory-guided movement (Houk et al., 2007). Posterior cortical areas (such as the lower visual system) seem to be less connected with the BG. Specificity is preserved throughout the projection target of the BG which are the 50-60 nuclei of the thalamus (Herrero, Barcia & Navarro, 2002) and which project back to specific areas of the cortex.

3. Theory of reinforcement learning

The canonical reinforcement learning network (Sutton & Barto, 1998) has an input layer on which the activity of exactly one unit codes the state s of the agent and an output layer on which the activity of one unit codes the action a the agent is going to choose given the input. Fig. 2. a) shows the architectures of two algorithm classes, TD-learning and SARSA. Input and output layers are termed state and actor in both implementations. A critic may be used only to guide learning. Given a random initial state (position) of the agent within the limited state space, and another state at which a reward r is consistently given, the agent learns to maneuver directly to the rewarded state.

In the case of TD-learning all states are assigned goodness values v that represent the sum of discounted future rewards and are kept by the critic in a lookup table V. A distant reward will be discounted in that it keeps only a proportion, e.g. $\gamma \approx 0.9$, of its original value for each step required to get it. So if the reward r will be reached in n steps, then the current state will be worth $v = r\,\gamma^n$.

Standard reinforcement learning lacks a "working memory" to backtrack recently visited states when a reward is given. Instead, a state value v is updated from the value v' of the neighbor state that is visited in the next step. Since then one step is done, the reward was further away in the previous state, hence $v = \gamma\, v'$. If the reward is given, instead $v = r + \gamma\, v'$. This equation will be inconsistent for neighbouring states during early learning. Step (5) of the algorithm in Fig. 2c) quantifies this error which is then used in steps (6) to update the value v of the previous state. The difference in time between the new estimate $r + \gamma\, v'$ and old estimate v taken in step (6) bestows this class of algorithms the name Temporal Difference (TD) learning.

The actor-critic architecture employs a dedicated neuron, the critic, to encode the expected future reward – or the values v – in its connections V. The connections Q to the actor which encode the action policy are separate[1]. The critic influences the actor update, step (7) in Fig. 2, through its prediction error δ. Vice versa, the current action policy determines which states the agent will visit next, and this feeds back into the update of the critic's value v.

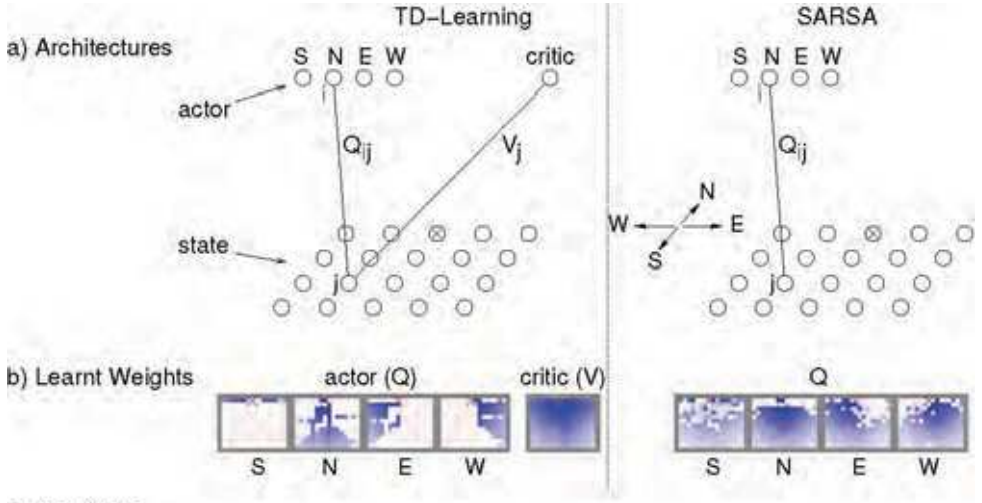
SARSA[2] encodes the value of state-action pairs (s, a) instead of the value of states. It may be implemented with a critic neuron that is connected to all state units and all action units. Fig. 2, right, shows an implementation without a critic, using only the weights from the state units to the action units. These store the state-action values Q and are also used to choose the action in step (3). However, computation of v and v' involves state units j, j′ and action units i, i′, hence, crosstalk involving some lateral connections must exist.

The state values V (or Q in case of SARSA) depend on the action strategy, because that influences the number of steps required to reach the reward. The action strategy is implemented in step (3) of the algorithm. Note the stochastic choice of actions. A deterministic agent may select a long path with a gradual increase of value rather than a short path on which it hasn't yet assigned any value to some states. The stochasticity allows for exploration of new states over exploitation of a previously thought optimal strategy. During learning, weights Q and hence the inputs h in step (3) become larger and so the character of action choice becomes more deterministic.

---

[1] These connections are sometimes called "P" to denote action preferences.

[2] SARSA computes the values from (s, a, r, s0, a0), hence the name.



c) Algorithms

	TD-Learning	SARSA
Initialize to new starting position		
(i1) Read old state	$s_j = 1$ if in state j, else $s_j = 0$	
(i2) Choose action	$\mathrm{Prob}(a_i = 1) = \frac{e^{2h_i}}{\sum_k e^{2h_k}}$ where $h_i = \sum_l Q_{il} s_l$	
(i3) Value	$v = \sum_l V_l s_l = V_j$	$v = \sum_{k,l} Q_{kl} a_k s_l = Q_{ij}$
Repeat until trial ends successfully (or out of valid area)		
(1) Act; Read new state	$s'_{j'} = 1$ if in state j', else $s'_{j'} = 0$	
(2) Read reward	r	
(3) Choose action	$\mathrm{Prob}(a'_{i'} = 1) = \frac{e^{2h_{i'}}}{\sum_k e^{2h_k}}$ where $h_{i'} = \sum_l Q_{i'l} s'_l$	
(4) Value	$v' = \sum_l V_l s'_l = V'_{j'}$	$v' = \sum_{k,l} Q_{kl} a'_k s'_l = Q'_{i'j'}$
(5) Prediction error	$\delta = r + \gamma v' - v$	
(6) Critic weight update	$\Delta V_j \propto \delta s_j$	—
(7) Actor weight update	$\Delta Q_{ij} \propto \delta a_i s_j$	
(8) New becomes old	$s \leftarrow s', a \leftarrow a', v \leftarrow v'$	

Fig. 2. TD-learning vs. SARSA. a) Architectures. The actor-critic architecture used in TD-learning has weights from all state space units to all action units and to the critic. SARSA is missing these critic weights, but there is additional information flow via links that are not shown. b) Trained weights for a toy problem. Dark blue denotes strong positive weights. The rewarded position is indicated by a "×" in the 16×12 state space. c) Algorithms. Actor-critic learning assigns a value to a state s, SARSA to a state-action pair (s, a). Note: (i) The value in step (4) reduces to a single weight because only state unit j′ and action unit i′ have activation 1, others are 0. (ii) In TD-learning, step (3) may be done after (7).

Relation to Basal Ganglia

The lateral inhibition in the striatum might ensure that neurons will be active only in a small focused region which directly represents the state, just like a single active neuron denotes the state in the models. In such a localist – as opposed to a distributed – code, a neuron does not participate in the coding of several completely different states. Thereby an assignment of reward to all active units will not interfere with other states, which is important in the critic-and actor update steps (6), (7) in Fig. 2. In accordance with this demand, the striatum is known as a 'silent structure', in which only a small percentage of the dominant neuron type, the spiny projection neurons, is strongly active at any one time[3].

If the striatum encodes the state s_j, and if the GPi/SNr encode actions a_i, then the δ could modulate learning by dopaminergic projections to either striatum or GPi/SNr neurons to form the multiplicative factor in the actor update, step (7).

Dopaminergic neurons are mainly found in the VTA and the SNc, and they do not have spatial or motor properties (Morris, Nevet, Arkadir, Vaadia & Bergman, 2006). Corresponding to the value δ, dopaminergic neurons exhibit bursts of activity in response to unexpected rewarding stimuli or conditioned stimuli associated with those rewards (Ungless et al., 2004). Their firing correlates with expected reward values, i.e. probability times magnitude of the reward (Tobler, Fiorillo & Schultz, 2005). While dopamine neurons generally respond briefly to unexpected reward delivery, trained neurons will respond briefly to the cue that predicts an upcoming reward, but not to the expected reward itself, and their baseline firing will be suppressed, when an expected reward fails to be delivered (Schultz, Dayan & Montague, 1997).

Biological Support for SARSA

When monkeys choose to reach one of two levers, one paired with a frequent reward and the other with a less frequent reward, they do not always choose the more frequently rewarded action, even if overtrained. Instead, they adopt "probability matching", a suboptimal strategy in which the distribution of responses is matched to the reward probabilities of the available reward. This allows neural activations to be measured, when deciding for the lesser rewarded action (Niv, Daw & Dayan, 2006; Morris et al., 2006). Within just 200 msec after stimulus presentation dopamine neurons fire in proportion to the reward associated with the lever that they will reach at later, even if the reach is performed seconds later. In particular, they will fire less if the monkey is going to choose the poor reward. These results seem to contradict the actor-critic models in which the value v of a state is independent of the next action[4]. They are in accordance with SARSA, in which the value depends on the state and the action that is chosen (but not yet executed).

Multiple Tasks During Learning

Rothkopf and Ballard (2007) hint at a problem that arises in realistic scenarios when multiple reinforcement strategies are learning concurrently. Since there is only one dopamine reward signal, not only the successful strategy, but all active strategies would receive it. Their solution is to share it: each strategy consumes an amount of the reward that is proportional to the reward that it expects from the current state transition, the

[3] Brown et al. (2004) suggest that feedforward inhibition causes such sparse firing, because recurrent (feedback) inhibition would require significant activation of many neurons to be effective.

[4] However, a predictive (cortical) input to the basal ganglia could denote already the next state.

difference of the corresponding values. Unfortunately, a strategy would not receive any amount of the reward if the reward comes completely unexpected under this strategy. This might be remedied by taking into account confidence values to each strategy's prediction. In any case, the different parallel loops need to communicate, possibly via the indirect pathway of the basal ganglia.

Exploration – Exploitation

Sridharan, Prashanth and Chakravarthy (2006) address the problem that a RL network has to produce randomness in some controllable fashion, in order to produce stochastic action choices. For this purpose they implement an oscillatory circuit via reciprocal excitatory-inhibitory connections. The indirect pathway (see Fig. 1) represents a suitable oscillatory circuit in that the STN excites the GPe and in turn receives inhibition. Together with short-range lateral inhibition the model produces chaotic oscillatory activity that becomes more regular only with stronger input from the cortex (Sridharan et al.,2006). They propose that, in case of weak or novel sensory input, the irregular firing causes an agent to behave more randomly and thereby to explore the state space. A biological manifestation of randomness could be in the pauses by which GPe neurons randomly and independently of each other interrupt their otherwise regular high-frequency firing (Elias et al., 2007). These pauses last approximately half a second and happen on average every 5 seconds with Poissonian interpause intervals. There are less pauses during high motor activity, indicating less randomness during performance.

4. Implementations

A central idea about 'lower' and 'higher' parts of the brain is that lower centers "swap out" functions that they cannot perform themselves. At the lowest level we might find distributed control circuits such as in some inner organs, as well as spinal cord reflex mechanisms. Since they function autonomously we may not actually cast them into a hierarchy with other brain structures.

The reticular formation at a very low level is mature at birth and regulates the choice of basic behaviours such as eat, fight or mate. As a centralized structure it can coordinate these behaviours, which the distributed control circuits would not be able to do (Prescott, 2007). Yet it lacks sophisticated learning capabilities and cannot cope with a complex and changing environment.

The basal ganglia implement a memory of successful actions in performing stimulus-response mappings that lead to rewards based on experiences of previous stimulus-response performance. Whether any reward is of interest may be set by a currently active basic behaviour (a thirsty animal will appreciate water but not food). So the reticular formation may have control over the basal ganglia, in selecting sub-circuits for different types of reward and strategies.

But the sensory stimuli from a complex environment are not necessarily suitable as a 'state' in reinforcement learning. A situation like "food is behind the door" is hardly represented suitably. Suitable state representations are unlikely to be learnt from reinforcement learning, and unsupervised learning is a better candidate. The basal ganglia may "swap out" such functionality to the cortex. To learn useful representations, unsupervised learning in the cortex may be guided by rewards and attentional selection (see Section 5).

The cortex features various functionalities despite its homogeneous structure. (i) Preprocessing in low hierarchical levels in the posterior cortex. The purpose is to transform light or sound into meaningful entities like objects, words or locations. (ii) Working memory in higher hierarchical levels in more anterior cortex. An example usage is for task setting: the strategy to use, or the reward to expect, is dependent on an initial stimulus that must be held in memory. The cortex may thereby determine which part of the basal ganglia to use, possibly overriding influence from the reticular formation. (iii) Motor primitives, presumably on a middle hierarchical level, in the motor cortex. For example, an action like "press the left lever" is a coordinated temporal sequence of muscle activations. Such action primitives reside in the motor cortex, and also the cerebellum, which we do not address further, is involved.

Architectural Choices

Tiered architectures are common in robotics. They allow to implement short-term reactive decisions while at the same time pursuing long-term goals. They also allow for computer programs to be implemented in modules with minimal inter-modular communication.

Brooks' subsumption architecture is an early example (Brooks, 1986). From the robot's lowest control layer to the highest layer, actions are for example: "avoid an object" — "wander around"—"explore the world"—"create a map", the latter of which may be the ultimate goal of a particular robotic application. The layers of such an architecture are however not directly identifiable with brain structures such as reticular formation — basal ganglia — cortex.

Unlike structured computer programs the 'modules' of the brain are highly inter-dependent. Computations involve multiple brain structures, and actions are often redundantly executed in parallel. For example saccades are destroyed by a combined lesion of the midbrain superior colliculus (SC) and the cortical frontal eye field (FEF), but not by a lesion of either of the two (Sparks, 2002). Another design principle of the brain is recurrence – connections form loops within and between brain structures.

Several models which mainly focus on the basal ganglia implement a larger loop structure.

The basic loop (Fig. 1) is Striatum → GPi/SNr → Thalamus → Cortex → Striatum. This loop is topographic in the sense that there are separate parallel loops, each for a specific feature, thought or action (Houk et al., 2007). Action on the level of the GPi/SNr activates the entire corresponding loop that includes slices of the thalamus and cortex as well.

Robot Action Selection

Prescott, Stafford and Gurney (2006) use a basal ganglia model for basic behaviour selection in a Khepera robot. The robot removes cylinders from its arena using five action patterns in the natural order: cylinder-seek, cylinder-pickup, wall-seek, wall-follow, cylinder-deposit. Scalar salience signals for each of these actions, which depend on perception and motivation, are the input to the basal ganglia. These are implemented with standard leaky integrator neurons and hand-set parameters to select coherent sequences of actions. A sophisticated embedding architecture complements the basal ganglia model: perceptual sub-systems (e.g. visual cortex) and motivational sub-systems (e.g. reticular formation) for computation of the salience signals; stereotyped, carefully timed "fixed action patterns" (e.g. motor cortex) for action execution. A "busy signal" prevents currently performed actions from being interrupted by other action bids. In the model of Brown et al. (2004), such a suppression of lingering actions is done via the STN which sends diffuse excitation to the inhibitory BG output nuclei GPi/SNr.

Working Memory Control

Hazy et al. (2007) generalize action selection to the selection of working memory representations in the pre-frontal cortex (PFC). This tackles the temporal credit assignment problem in trace conditioning where there is a gap between the conditioned stimulus and the reward. The working memory capacity of the PFC bridges this gap and delivers sustained input to the basal ganglia. Working memories with different time spans in parallel loops allow for the execution of nested tasks. Their example application is the 1-2-AX task, in which a subject after seeing a '1' must identify the consecutive letters 'A-X', but after seeing a '2' must identify the sequence 'B-Y'. The numbers '1', '2' are memorized for a longer duration in an 'outer' loop. An 'inner' loop identifies the desired letter sequence within a short duration. A third loop elicits the motor response. While the basal ganglia resolve only these loops, the much larger cortex distinguishes also the contents within the loops. In the 1-2-AX task these are the values of the numbers and the digits. The model PFC stores them in hypercolumn-like "stripes" with one of several entries in a stripe being active in a winner-take-all fashion. Gating is nevertheless accomplished in the basal ganglia that does not need to reflect the individual features within a stripe.

Basal Ganglia Mediate Cortical Control on Superior Colliculus

The superior colliculus (SC) is a phylogenetically old structure eliciting reactive saccades from direct retinal input. Planned saccades are elicited on the SC only via direct cortical input and concurrent disinhibition by the basal ganglia. This suggests that planned saccades are driven by expected reward (Hikosaka, 2007). Brown et al. (2004) implement such an extensive circuit and simulate saccade tasks which involve target selection and timing.

Their model takes into account the cortical layer structure. 'Planning' cells in cortical layer 3 with sustained activity send preparatory bids to the basal ganglia, while associated 'executive' cells in layer 5 generate phasic outputs if and when their basal ganglia gate opens. Planning cells are modulated by layer 6 cells which possibly reside in higher-level cortical area (PFC) that is in control. These same cells of cortical layer 6 are a source of excitation to the thalamic cells whose disinhibition allows plans to execute.

A hypothesis of Brown et al. (2004) is that thalamo-striatal connections (not shown in Fig. 1) become active in trials during which premature release of a movement leads to non-reward; this shall lead to a learned activation of the indirect channel and therefore guide the learning of 'STOP' responses.

Basal Ganglia Instruct Cortex

There appear to be two positions related to the association between the prefrontal cortex (PFC) and basal ganglia. The conventional view is that the PFC drives the learning of the basal ganglia. This is mainly based on the fact that the striatum neurons require numerous synchronous inputs from cortex (and thalamus) to become active. An alternative view is that while the dopamine system 'teaches' the striatum, the basal ganglia teaches the cortex through the basal ganglia-thalamo-cortical loop (Laubach, 2005; Graybiel, 2005). Our model of Weber et al. (2006) utilises this alternative.

Areas of the motor cortex execute action primitives; on the other hand, the basal ganglia are well equipped for learning these actions by reinforcement learning in the first place. In our model an action that has been acquired by the basal ganglia is then imitated by the motor cortex. Thereby the resources used for reinforcement learning, such as the large state space that may reside in the striatum, would be available for further learning. See Section 5 for a description of this visually guided robot docking action (Weber, Wermter & Zochios, 2004). Both of these levels of neural processing have been implemented on a PeopleBot robot.

In our model of Weber et al. (2006) the motor cortex reads the visual input and motor output of the basal ganglia. It establishes an internal representation of these input-output pairs by unsupervised self-organization (Hinton, Dayan, Frey & Neal, 1995). With 'incomplete' input in which vision is present but the action missing, the network will find a 'complete' internal code from which it will generate an appropriate action. Horizontal hetero-associator weights on the internal layer associate the current representation with a future representation one time step ahead, and thereby perform prediction, allowing for mental simulation of an action.

Experimental evidence supports our model. During associative learning, Pasupathy and Miller (2005) found earlier changes of neural activity in the striatum than in the PFC. In their study, primates were rewarded if they made saccades to a certain direction, dependent on the appearance of a complicated cue shown at the fixation point. The primate learnt the rewarded direction by trial and error. Once the relationships were learned the input-response pairs where reversed. When relearning the appropriate behaviour to the input, the striatum was found to have direction-specific firing almost straight away. In contrast, the PFC only gained direction selectivity following 15 correctly performed trials. This is consistent with the striatum training the PFC.

Jog, Kubota, Connolly, Hillegaart and Graybiel (1999) trained rats to make a left-right decision in a T-maze task. Striatal neurons which were initially active at the point of the junction became less active when the task had been learnt. Instead, they increased their activities at the beginning and at the end of the task. This suggests that the striatum might be putting together sequences of known behaviours that, once learned, are executed elsewhere.

Task Switching

Representing actions on the cortex might also make them easier to control by other cortical areas. Prelimbic and infralimbic regions of rat prefrontal cortex were shown to remember different strategies and aid in switching between learnt strategies (Rich & Shapiro, 2007). In an extension of our model of the motor cortex (Weber et al., 2006) we therefore showed that language input to another cortical area can influence which motor sequence on the motor cortex representation to recall (Wermter, Weber, Elshaw, Gallese & Pulvermüller, 2005). These cortical learning principles can lead to language-guided neural robots in the future.

5. Visual system

The actor-critic model of reinforcement learning has been used to perform various robot actions, such as camera-guided robot docking (Martínez-Marín & Duckett, 2005). In our approach (Weber et al., 2004), we first trained the peripheral vision so that it can supply a visually obtained state as input to the action selection network.

Overall, there are three processing steps and associated training phases involved in the learning of the docking behaviour (see Fig. 3). First, training the weights between the visual input and the "what" area by unsupervised learning. The learning paradigm is that of a generative model (hence the feedback connections in Fig. 3) in which the color image is reconstructed from sparse activations of neurons in the "what" area. Second, training the lateral weights within and between the "what" and the "where" areas by supervised learning. For this purpose, a supervisor placed a blob of activation onto the position on the "where" area which corresponded to the correct position of the object within the image. After learning, an attractor network covering the "what" and the "where" areas creates the

"where"representation by pattern completion if only the "what" representation is supplied as input.

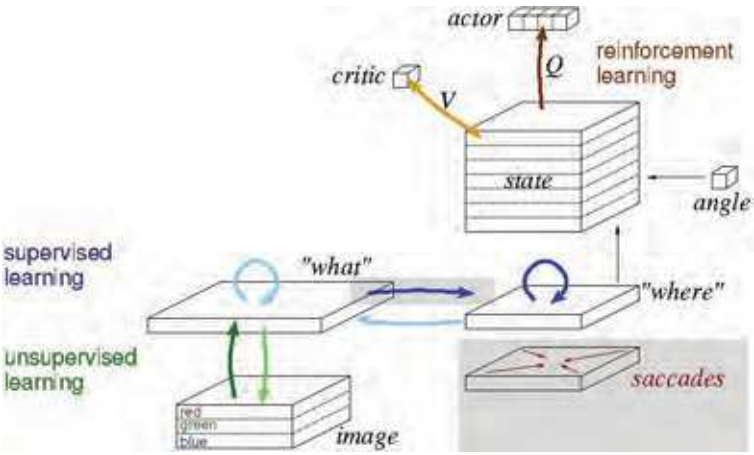


Fig. 3. Neural architecture for visual pre-processing and reinforcement-learnt action. Thick arrows denote trained weights. Only the ones depicted dark are used during performance while those depicted bright are involved in training. The area on shaded background labelled 'saccades' is assumed to perform saccades to bring the object from any location on the 'where' area to its center, as indicated by the arrows pointing to the middle. Saccades can be used here to replace supervised learning of the "what" → "where" connections (shaded background) by reinforcement learning (see Section 5.1).

The robot needed to approach the table at a right angle. For the final step therefore the visual "where" representation of the object was augmented by the robot angle w.r.t. the table, here discretized into just seven angle values. This outer product yielded the state space, a 3-dimensional block in which one unit's activity denoted the visual object position and the robot angle; in other words, seven layers of the visual space, one for every possible robot angle. Finally, the weights from this state space to the critic and the four actor units, denoting 'forward', 'backward', 'turn left' and 'turn right', were trained by TD-learning.

The critic weights assign each state a value v which is initially zero. For each trial the robot is initialized to a random starting position and the steps of Fig. 2 are followed until the reward is obtained. The reward signal is given when the target is perceived in the middle of the lower edge of the visual field (hence at the grippers) and when the robot rotation angle is zero. When the robot hits the table or looses the object out of sight, then a new trial is started without a reward. During learning, states that lead quickly to the goal will be assigned a higher value v by strengthening their connections V to the critic unit. The weights Q to the motor units which have been activated simultaneously are also increased, if the corresponding action leads to a better state.

In short, we have used a simple 'what-where' visual system as a preprocessing module, supplying suitable input to a state space for reinforcement learning. In-line with the classical

view, this simplified visual system learns from visual data irrespective of the use of vision for action and reward-seeking.

5.1 Reward in the visual system

In recent years evidence is accumulating that even the lower visual system such as the primary visual cortex V1 is sensitive to reward learning. Shuler and Bear (2006) repeatedly presented rats a light flash to one eye followed by a reward given after two seconds. V1 neurons acquired reward-dependent responses such as sustained responses after visual stimulus offset, or increasing responses until the time of the (expected) reward. Schoups, Vogels, Qian and Orban (2001) trained monkeys to discriminate oriented bars (such as distinguishing 45° from 43° orientations), after the presentation of which they had to respond with saccades to a certain direction to receive a juice reward. After training, the slopes of the orientation tuning curves were increased in V1 neurons tuned to orientations near the trained orientation[5]. On the other hand, no modifications of the tuning curves were observed for orientations that had been shown as often but which were not decision relevant.

But learning in the adult visual system is not always reward dependent. Furmanski, Schluppeck and Engel (2004) trained subjects to detect very low-contrast oriented patterns, following which they indicated a decision, but which did not incur a reward. This fMRI study revealed increased V1 responses for practiced orientations relative to control orientations. However, Vessel (2004) conjectures that stimuli that make sense and are richly interpretable on a higher level are 'rewarding' and perceived as pleasurable. He recalls that there is an increasing number of opiate receptors as one traverses up the visual hierarchy. Hence, mere neuronal activation might be regarded as reward and be utilized in learning algorithms.

Saccade Learning

In Weber and Triesch (2006) we have trained saccades using a reward signal made only from visually-induced activation. The model exploits the fact that the fovea (the center of the retina) is over-represented in visual areas. Saccades to an object are rewarded dependent on the resolution increase of the object — a value that is higher the closer the object is brought to the fovea. Motor units which code for a certain saccade length, and which become active in a noisy competition, compete via limited afferent connections. A motor unit that brings the object closest to the fovea will learn with the highest reward modulation and ultimately win. Since there is evidence for a different learning mechanism for horizontal saccades, we applied this algorithm for the learning of vertical saccades in combination with a different algorithm for horizontal saccades.

When saccades have been learnt, we can assume that neurons in higher visual areas of the "where" pathway exist which code for saccades of a certain direction and amplitude, as indicated in the shaded area of Fig. 3. These are then akin to action units. With the algorithm of Weber and Triesch (2006) we can then learn the "what" → "where" connections by reward-based learning instead of by supervised learning.

---

[5] Neurons which adapted their tuning curves were found only in supra- and infragranular layers of V1 where there are dense intra- and inter-area horizontal connections as well as inter-area top-down connections. Neurons in layer IV which receive bottom-up input from retina/thalamus did not adapt.

Gaze Following

The potential of a purely visual stimulus as a reward is also used in a RL model of how infants learn to follow the gaze of the mother (Triesch et al., 2007), a skill which infants learn only after 18 months of age. The model assumes an infant's tendency to look frequently at the mother's face. It assumes further that the mother then looks to the left or the right, and that there is an interesting (rewarding) stimulus where the mother looks. The infant initially cannot make use of the mother's gaze direction, but after making (initially random) sample eye movements, it will find out that rewarding stimuli can be found in the line of sight of the mother. The model predicts a mirror-neuron like premotor representation with neurons that become activated either when the infant plans to look at a certain location or when the infant sees the mother looking in the direction of that location.

5.2 Attention-gated reinforcement learning

Attention-Gated Reinforcement Learning (AGREL) (Roelfsema & Ooyen, 2005) is a link between supervised and reinforcement learning for 1-of-n classification tasks. In supervised learning of such tasks the teacher's learning signal is 1 for the correct output unit and 0 for the other output units, and is given for every data point. The rules of reinforcement learning are that if the network – of which the output will be stochastic winner-take-all – guesses correctly, then a reward signal is given, else not. AGREL gives learning rules which in this case lead to the same average weight changes as supervised backpropagation learning, albeit learning is slower due to insufficient feedback when the network guesses incorrectly.

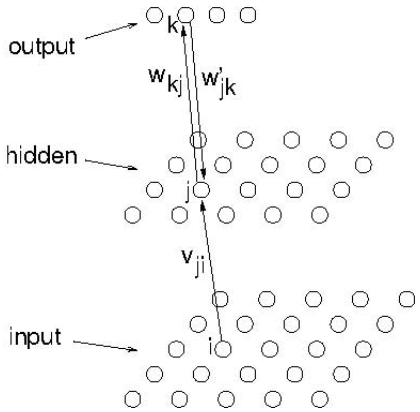


Fig. 4. Architecture of AGREL. One-in-n of the output units is active at a time, just like in TD-learning and SARSA. The input and hidden layers, however, may have a distributed code.

The AGREL architecture is that of a multilayer perceptron and shown in Fig. 4. An input unit i conveys its activation x_i via weight v_{ji} to the activation y_j of hidden layer unit j which has a logistic transfer function

$$y_j = \frac{1}{1+\exp(-h_j)} \quad \text{with} \quad h_j = \sum_i v_{ji} x_i$$

Activations are then conveyed to an output unit k, while the output units compete via a softmax function:

$$e_k := \text{Prob}(z_k = 1) = \frac{\exp(h_k)}{\sum_{k'} \exp(h_{k'})} \quad \text{with} \quad h_k = \sum_j w_{kj} y_j .$$

The actual, binary output z_k of the neuron follows the probability e_k of being active, which is thus the average activation.

Learning

In order to understand AGREL, lets first consider error backpropagation[6]. The average update of a weight w_{kj} from a hidden unit j to an output unit k is:

$$E(\Delta w_{kj}) \propto (t_k - e_k)\, y_j , \tag{1}$$

where t_k is the teacher signal and e_k is, in backpropagation, the continuous activation on output neuron k. Unlike supervised backpropagation, AGREL considers only the winning output neuron, $k = s$, for learning. Now we apply Eq. 1 for reinforcement learning in which we distinguish two cases, unrewarded and rewarded trials. For unrewarded trials, which means $t_s = 0$, Eq. 1 becomes

$$E(\Delta w_{sj}) \propto -e_s\, y_j =: f(\delta)\, e_s\, y_j ,$$

and for rewarded trials, where $t_s = 1$, Eq. 1 becomes (defining $\delta := t_s - e_s$)

$$E(\Delta w_{sj}) \propto \underbrace{(1 - e_s)}_{\delta}\, y_j = \delta\, y_j \underbrace{\frac{e_s}{1-\delta}}_{1} =: f(\delta)\, e_s\, y_j .$$

In order to make both weight update steps consistent, one defines

$f(\delta) := -1$ for unrewarded trials, and $f(\delta) := \frac{\delta}{1-\delta}$ for rewarded trials.

To complete our brief treatment of AGREL,the top-down feedback weights w_{jk} to the hidden layer learn with the same rule as the w_{kj}. A weight v_{ji} from an input unit i to a hidden layer unit j is updated according to:

$$\Delta v_{ji} \propto x_i\, y_j\, (1 - y_j)\, w'_{js}\, f(\delta).$$

Hence, learning of the weights from the input to hidden unit j scales with the weight w_{js} that this unit receives from the only active output unit s. The term $f(\delta)$ depends on whether the winning unit equals the correct output.

Applications

AGREL works only with immediate rewards and does not build up action strategies as TD-learning does. While TD-learning requires that only one input unit is active at a time, AGREL accepts a distributed code as input. Applications are therefore in the sensory system where a classification of a stimulus needs to be made, and where some kind of reward signal is promptly available.

For example, monkeys had previously been trained to categorize faces, and it emerged that neurons in the inferotemporal cortex preferentially encode diagnostic features of these faces, i.e. features by which categories can be distinguished, as opposed to features that vary irrespective of categories. Roelfsema and Ooyen (2005) showed that neurons of the hidden layer

[6] Here we outline chapter "4.1 Average Weight Changes in AGREL" of Roelfsema and Ooyen (2005) from back to front.

in AGREL also learn preferentially to code for diagnostic features. This explains that feature representations in the sensory cortex are not merely dependent on the (statistics of) sensory input, but are tailored to decisions, and thereby aimed at rewards.
Another example is the simulation of the abovementioned experiments of Schoups et al. (2001): orientation selective neurons in V1 (AGREL's hidden layer) adjust their tuning curves to distinguish orientations that are being classified, while neurons that are as often activated, but without action relevance, do not adjust their tuning curves.
In Franz and Triesch (2007), vergence eye movements were trained using AGREL so to fixate an object with both eyes in depth. No external reward was given, but a successful focusing of both eyes on a single point in space led to binocular zero-disparity cells having a particularly high activation, which was regarded as a reward. The model predicts a variety of disparity tuning curves observed in the visual cortex and thereby presents additional evidence that purely visually-induced activation can constitute a reward signal.

6. Beyond behaviourism

Reinforcement learning (RL) theory has probably been the most influential outcome of behaviourist psychology, and neuroscientific support for it continues to grow. It seems very likely that there is RL in the basal ganglia, complemented by unsupervised learning in the cortex and supervised learning in the cerebellum (Doya, 1999), possibly building upon genetic 'learning' in the reticular formation.
However, the brain still resists a unifying description. Baum (2004) argues that either cognitive behaviour can be described in a compressed way using a few theories like RL. Or it has been reasonably well optimized by evolution, and theories which are simple enough to comprehend inevitably miss out a lot of significant detail. In attempting to uncover simple psychological principles, the behaviourists have left a legacy in which the same theories which illuminate brain activity, can render us blind to other significant aspects.

6.1 Experimental conditions

In the early 1970's, every psychology department still had a "rat lab". The researchers did what they could to control the experimental conditions. But a rat swimming in a water maze is still aware that light glistens off the water in subtly different ways according to its direction, and although the walls were white and high, the sounds of birdsong outside the window or footsteps in the corridor were still there, providing good orientation cues. While human beings are often only vaguely aware of their surroundings, animals are highly attuned to their environment. A neuroscientist recounts how simply wearing a different lab coat can radically change an animal's behaviour (Panksepp,1998, p.18).
Also, the fashion at the time was to write up experiments on living animals in the same formal manner that has proven so useful when dealing with non-living subject matter. Conceptual analysis, a view of the world in stimulus-response terms, left no place for context (either external or internal). An observer at the time might see that the written accounts of the experiments were patently not portraying what was happening. They were merely reporting how the researchers interpreted what they saw, suppressing often more interesting behaviours as "irrelevant". The purpose of the experimental conditions was to

effectively deprive the animals of every possible natural behaviour apart from the sought-for response. Much of the time, the animals would do anything except the behaviour under test.
The first legacy of the behaviourists has been an oversimplified account of the experimental conditions under which RL was investigated. We have attempted to demonstrate that the maths of RL can be formulated under assumptions that are also supported by the behaviour of animals in their natural environment. Oversimplification can be avoided with a careful, honest eye on neuroscientific results and the use of robots to test the theories in a practical context.

6.2 Embedded behaviours

The founder of ecology, Konrad Lorenz, identifies adaptation (including RL) as merely one of nine types of cognitive behaviour (Lorenz, 1996). He claims RL is a phylogenetically significant information acquiring system, but one which requires sophisticated subsystems for its operation. For example, *both* stimulus recognition and adaptive modifiability must be attuned to the message of success or failure coming from the activities terminating the whole action, which also must be capable of appraising significance. Mechanisms which enable the organism to distinguish reliably between biological success and failure are rarely as simple as the binary or scalar values used by RL.
Innate behaviours such as eat, fight and flee[7] have been described as mutually incompatible modes of vertebrate behaviour (Kilmer et al., 1969), and the candidate brain system for selecting between these kinds of action is the reticular formation in the brainstem. Without a cortex (and basal ganglia), electrical stimulation in the brainstem can induce complex and coordinated behaviours, including eating, grooming and attack (Berntson & Micco, 1976). These behaviours are modifiable in complex ways. Very little is understood concerning emotion or motivation, yet these are clearly crucial to a full understanding of RL in animals. Lorenz's cognitive behaviours include exploratory behaviour. This requires coherent activity concerning something which has not been learned, by definition. Yet it also has a rationale and a logic which is adaptive, distinguishing it from the blind randomness of behaviourists' descriptions, and of the standard RL protocols.
But just as the study of adaptation has grown significantly since the behaviourists' first formulations, so has RL. For some time, neuroscientific evidence has implicated the basal ganglia in RL. Panksepp (1998) (ch.8) revisits the literature on self-stimulation reinforcement and concludes that dopamine activity does not reinforce *consumatory* but *anticipatory* behaviours. So it is more akin to "the joy of the hunt". If this is the case, the old view of a specific stimulus becoming linked to a response via some general reinforcer seems unlikely. A better interpretation is that a stimulus set (which includes what is relevant in the wide-ranging context) is linked to the response via a reinforcer that is appropriate given that context
Hence, the second legacy of the behaviourists has been to encourage the widening of the scope of behaviours we study. It is inadequate to describe primitive behaviours as merely "innate" as this fails to account for the variety of their expression. Likewise, RL is not a

[7] Kilmer, McCulloch and Blum (1969) list the following: sleep, eat, drink, fight, flee, hunt, search / explore, urinate, defecate, groom, mate, give birth, mother the young, build a nest, and special speciesdependent forms of behaviour such as migrate, hibernate, gnaw, and hoard.

single, monolithic mechanism. What counts as a "stimulus" can range from a single neuron's activation to widely distributed patterns. Dopamine is unlikely to be the only reinforcer and the "response" can be as varied as any animal behaviour.

6.3 Neuroconstructivism

Lorenz has also criticized the assumption that the human mind, before any experience, was a tabula rasa, and the equivalent assumption that "learning" must "enter into" any physiological behaviour process whatever. The most common response to this is to claim that anything not learned must be innate. But this is an artificially narrow choice which follows from the philosophical assumptions science has inherited from Plato and Descartes. Science proceeds on the assumption that the only things that count are ideas which can be considered independent of anything else: objects which can be observed. This has served us well for centuries, and will continue to do so. But as psychology has already found out, studying cognitive behaviour in the same way leads to a number of difficulties. Inevitably, this will also become a problem for RL too, at some point.

Fortunately, there is an alternative viewpoint which promises to avoid many of the problems inherited from Cartesianism. Rather than assuming that things can be "atomic" as Plato suggested, Heidegger (1927/1962) emphasizes that all behaviour is executed in some context. We are thrust into a rich, pre-existing world and are actively living out our purposes in it from the start. There is no such thing as an object which has no context. Attempts to isolate things like "stimuli" and "responses" involve very high-level, sophisticated abstractions which Heidegger called *present-at-hand*, that is, we can examine them.

The neuroconstructivism of Mareschal et al. (2007) is typical of this more modern approach. They still expect all science to rest upon processes in the physical world, but this is in terms of a "coherent account". Components are intelligible as contributory constituents of the whole which gives them meaning. The system in turn is defined in terms of its components and their mutual relationships. One advantage of this formulation is that it simultaneously avoids the behaviourists' narrowness and the equally beguiling trap of modularity[8]. It is simply inappropriate in the real world to consider a "stimulus" as a single entity. Their conceptualization of "response" is equally sophisticated. According to neuroconstructivism, the outcome of almost every event is a distributed set of *partial representations* which are inevitably *context dependent*. All living systems (including cells) are considered *proactive* in the sense that they can be seen to be "active on their own behalf". This leads to an *interactive* interdependence between components, characterized by processes of cooperation and competition.

[8] Mareschal et al. cite Marr (1982) as their straw man here. According to them, Marr distinguishes independent computational, algorithmic and implementational levels. "For example, the same algorithm can be implemented in different technologies, or the same goals can be reached via different representational formats and transformations. The implication is that one can study and understand cognitive information processing without reference to the substrate in which it is implemented." Mareschal et al. (2007) (p.209) radically reject this view as an intelligent system must function in real time. Any sub-task is constrained not only by its functional definition but also by how it works, as it mustn't take too long. The implementation level cannot therefore be independent of the algorithmic.

6.4 Perception is an active process

Constructivists like Piaget (1953); Glasersfeld (1995) and Bickhard (2000) have emphasized that perception is essentially an *active* process. Psychological and physiological evidence (Gibson, 1979; Jeannerod, 1997; Noë, 2004) seems to indicate this is a viable theory. Although the basal ganglia are closely linked to action selection, there is a strong link with attention as well (Fielding, Georgiou-Karistianis & White, 2006). It is natural for researchers to focus on the most visible aspect of selected behaviour, the movement. But an action which has been selected is also being attended to. The implication of basal ganglia deficiencies in Attention Deficit Hyperactivity Disorder, following Teicher et al. (2000), confirms this compound nature of attention and action, as does the equitable treatment of sensory and motor basal ganglia afferents.

The model of Brown et al. (2004) illustrates this broader view of RL. It approximates the thousands of millions of interconnecting neurons in a model of less than 100 units, and tackles the complexity of the brain in a modular architecture[9]. It is sufficiently complex to take motivation and attention into account, as well as "learning". Indeed, the ability of the basal ganglia model to select between competing afferents may well provide a basis for choice – that element which distinguishes psychological learning from mere adaptation. In their words, "The basal ganglia interact with the laminar circuits in the frontal cortex and the superior colliculus to help satisfy the staging requirements of conditional voluntary behaviour." In the process they demonstrate that RL establishes stimulus control over *plans*, not responses, and provide a coherent alternative model of working memory.

The complexity of the basal ganglia, and their sensitivity to *context*, suggest a broadening of the simple stimulus-response view. Previous research becomes a special case. Stimuli become more natural and responses can be more than some action, as perception and attention are implicated in basal ganglia processing too.

The Heideggerian view that context is primary and conceptual data[10] is derivative, finds newfound support here and opens new possibilities. Many of the difficulties faced by Artificial Intelligence are the direct result of the Cartesian viewpoint that all context must be constructed from nothing. The recent success of embodied-embedded robotics research supports Heidegger's proposal that context is "given" (Wheeler, 2005). We have indicated above that the basal ganglia architecture seems to especially facilitate the processing of context alongside RL.

The understanding of RL has also widened. The change from a generally applicable pleasure-or-pain (with no clear cognitive implication) to a much more specific "sought-for" success (the cognitive link predicted by Interactivism (Bickhard, 1999)) means that RL is poised to address specific instances in a realistic way. RL, therefore, has much more in

---

[9] The assumption of modularity helps us conceptualize what is going on and formulate testable hypotheses, but it must be borne in mind that modularity is a function of our worldview, supported by its success in fields like computing and business systems. Writers like Braitenberg and Schüz (1998) go to great lengths to convey the messiness of the cortex. Overviews like Shepherd (2004) and Kandel, Schwartz and Jessell (2000) always indicate that, while neural pathways are a convenient way of getting to grips with the material, there are always exceptions and complications. Modularity is more an artefact of our scientific understanding than it is an aspect of the subject matter being explored.

[10] Heidegger would class this as "present-at-hand" – the stuff of scientific theories, or the disembodied "ideas" of Plato. Also the Cartesian view that things may be conceived of in isolation (that things-inthemselves are primary) is undermined by the same neuroscientific evidence.

common with the natural world and the variety of animal behaviour indicated by Lorenz (1996) than is warranted by the behaviourist evidence alone. This wider view places RL alongside other modern developments in philosophy and robotics. Such a combination must surely be grounds for hope that we will continue to see more robust and successful developments in artificial intelligence.

7. Acknowledgements

This work has been funded partially by the EU project MirrorBot, grant IST-2001-35282, and NEST-043374 coordinated by SW. CW, JT and AF are supported by the Hertie Foundation, and the EU projects PLICON, grant MEXT-CT-2006-042484, and Daisy, grant FP6-2005-015803. Urs Bergmann provided feedback on the manuscript.

8. References

Bar-Gad, I.; Havazelet-Heimer, G.; Goldberg, J.; Ruppin, E. & Bergman, H. (2000). Reinforcement-driven dimensionality reduction–a model for information processing in the basal ganglia. *J Basic Clin Physiol Pharmacol*, *11*(4), 305-20.

Baum, E. (2004). *What is thought?* MIT Press / Bradford. Berntson, G. & Micco, D. (1976). Organization of brainstem behavioral systems. *Brain Research Bulletin*, *1* (5), 471-83.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology*, *9* (4), 435-458.

Bickhard, M. H. (2000). Motivation and emotion: An interactive process model. InR. D. Ellis& N. Newton (Eds.), *The cauldron of consciousness: Motivation, affect and self-organization* (pp. 161–178). John Benjamins, Amsterdam.

Bloch, V. & Laroche, S. (1985). Enhancement of long-term potentiation in the rat dentate gyrus by post-trial stimulation of the reticular formation. *J Physiol*, *360*, 215-31.

Bogacz, R.& Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neur Comp*, *19*, 442-77.

Braitenberg, V. & Schüz, A. (1998). *Cortex: Statistics and geometry of neuronal connectivity* (2nd ed.). Springer Verlag.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEEJ Robotics and Automation*, *RA-2*, 14-23.

Brown, J.; Bullock, D. & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*, 471–510.

Dommett, E.; Coizet, V.; Blaha, C.; Martindale, J. & Lefebvre, V. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, *307*(5714), 1476-9.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*, 961-74.

Elias, S.; Joshua, M.; Goldberg, J.; Heimer, G.; Arkadir, D.; Morris, G. et al. (2007). Statistical properties of pauses of the high-frequency discharge neurons in the external segment of the globus pallidus. *J Neurosci*, *27*(10), 2525-38.

Fielding, J.; Georgiou-Karistianis, N. & White, O. (2006). The role of the basal ganglia in the control of automatic visuospatial attention. *Journal of the International Neuropsychological Society*, *12*, 657–667.

Fiorillo, C.; Tobler, P. & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898-902.

Fischer, T. (2003). *Charakterisierung des dopaminergen Systems bei transgenen Ratten mit einem Antisensekonstrukt gegen die m-RNA der Tryptophanhydroxylase*. Unpublished doctoral dissertation, Humboldt-University, Berlin.

Flaherty, A. W. & Graybiel, A. M. (1991). Corticostriatal transformations in the primate somatosensory system. projections from physiologically mapped body-part representations. *Journal of Neurophysiology*, *66*, 1249–1263.

Franz, A. & Triesch, J. (2007). Emergence of disparity tuning during the development of vergence eye movements. In *Proceedings of the 6th IEEE International Conference on Development and Learning*.

Furmanski, C.; Schluppeck, D. & Engel, S. (2004). Learning strengthens the response of primary visual cortex to simple patterns. *Curr Biol*, *14*, 573-8.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin, Boston, MA.

Glasersfeld, E. von. (1995). *Radical constructivism: A way of knowing and learning*. London: Falmer Press.

Graybiel, A. (2005). The basal ganglia: learning new tricks and loving it. *Curr Opinion in Neurobiol*, *15*, 638-44.

Hazy, T.; Frank, M. & O'Reilly, R. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Phil.Trans. R. Soc.B*.

Heidegger, M. (1927/1962). *Being and time*. SCM Press, London. (Originally published as Sein und Zeit, Gesamtausgabe Volume 2; translated by John MacQuarrie and Edward Robinson)

Herrero, M.; Barcia, C. & Navarro, J. (2002). Functional anatomy of thalamus and basal ganglia. *Child's Nerv Syst*, *18*, 386-404.

Hikosaka, O. (2007). Basal ganglia mechanisms of reward-oriented eye movement. *Ann N.Y. Acad Sci*, *1104*, 229-49.

Hinton, G.E.; Dayan, P.; Frey, B. J. & Neal, R. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, *268*, 1158-1161.

Houk, J.; Bastianen, C.; Fansler, D.; Fishbach, A.; Fraser, D.; Reber, P. et al. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Phil.Trans.R. Soc.B*.

Humphries, M.; Gurney, K. & Prescott, T. (2005). Is there an integrative center in the vertebrate brain-stem? A robotic evaluation of a model of the reticular formation viewed as an action selection device. *Adaptive Behavior*, *13*(2), 97-113.

Jeannerod, M. (1997). *The cognitive neuroscience of action*. Blackwell, Oxford.

Jog, M.; Kubota, Y.; Connolly, C.; Hillegaart, V. & Graybiel, A. (1999). Building neural representations of habits. *Science*, *286*, 1745-9.

Kandel, E. R.; Schwartz, J. H. & Jessell, T. M. (2000). *Principles of neural science*. McGraw-Hill.

Kilmer, W. (1997). A command computer for complex autonomous systems. *Neurocomputing*, *17*, 47-59.

Kilmer, W.; McCulloch, W. & Blum, J. (1969). A model of the vertebrate central command system. *International Journal of Man-Machine Studies*, *1*, 279-309.

Laubach, M. (2005). Who's on first? What's on second? The time course of learning in corticostriatal systems. *Trends in Neurosci*, *28*(10), 509-11.

Lewis, B. & O'Donnell, P. (2000).Ventral tegmental area afferents to the prefrontal cortex maintain membrane potential 'up' states in pyramidal neurons via D1 dopamine receptors. *Cerebral Cortex*, *10*(12), 1168-75.

Lingenhöhl, K. & Friauf, E. (2004). Giant neurons in the caudal pontine reticular formation receive short latency acoustic input: An intracellular recording and HRP-study in the rat. *J Comparative Neurology*, *325*(4), 473-92.

Lorenz, K. (1996). Innate bases of learning. InK.H. Pribram & J.King (Eds.), *Learning as self organization* (pp. 1–56). Lawrence Erlbaum.

Mareschal, D.; Johnson, M.; Sirois, S.; Spratling, M.; Thomas, M.& Westermann, G. (2007). *Neuroconstructivism: Perspectives and prospectives (volume i)*. Oxford University Press.

Marr, D. (1982). *Vision*. Freeman, San Francisco.

Martínez-Marín, T. & Duckett,T. (2005). Fast reinforcement learning for vision-guided mobile robots. In *Proc IEEE International Conference on Robotics and Automation (ICRA 2005).*

Morris, G.; Nevet, A.; Arkadir, D.; Vaadia, E. & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neurosci*, *9*(8), 1057-63.

Niv,Y.; Daw, N. & Dayan, P. (2006). Choice values. *Nature Neurosci*, *9*(8), 987-8.

Noë, A. (2004). *Action in perception*. MIT Press.

Panksepp, J. (1998). *Affective neuroscience*. New York: Oxford University Press.

Pasupathy, A. & Miller, E. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, *433*, 873-6.

Piaget, J. (1953). *The origin of intelligence in the child*. Routledge and Kegan Paul.

Prescott, T. (2007). Forced moves or good tricks in design space? Landmarks in the evolution of neural mechanisms for action selection. *Adaptive Behavior*, (to appear).

Prescott, T.; Stafford, T. & Gurney, K. (2006).A robot model of the basal ganglia: Behavior and intrinsic processing. *Neural Networks*, *19*, 31-61.

Reynolds, J.; Hyland, B. & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67-70.

Rich, E. & Shapiro, M. (2007). Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. *JNeurosci*, *27*(17), 4747-55.

Roelfsema, P. & Ooyen, A. van. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neur Comp*, *17*, 2176-214.

Roesch, M. & Olson, C. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. *Science*, *304*(5668), 307-10.

Rothkopf, C.& Ballard, D. (2007). Credit assignment with Bayesian reward estimation. In *COSYNE -Computational and Systems Neuroscience.*

Samejima, K.; Ueda, Y.; Doya, K. & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337-40.

Schoups, A.; Vogels, R.; Qian, N.& Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*, 549-53.

Schultz,W.; Dayan, P. & Montague,P. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-9.

Shepherd, G. M. (Ed.). (2004). *The synaptic organization of the brain*. Oxford University Press.

Shuler, M. & Bear, M. (2006). Reward timing in the primary visual cortex. *Science*, *311*, 1606-9.

Sparks, D. (2002). The brainstem control of saccadic eye movements. *Nat Rev Neurosci*, *3*, 952-64.

Sridharan, D.; Prashanth, P. & Chakravarthy, V. (2006). The role of the basal ganglia in ex ploration in a neural model based on reinforcement learning. *Int J Neur Syst*, *16*(2), 111-24.

Stafford, T. & Gurney, K. (2007). Biologically constrained action selection improves cognitive control in a model of the stroop task. *Phil.Trans.R. Soc.B*.

Sun, X.; Zhao, Y. & Wolf, M. (2005). Dopamine receptor stimulation modulates AMPA receptor synaptic insertion in prefrontal cortex neurons. *J Neurosci*, *25*(32), 7342-51.

Sutton, R. & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Teicher, M.H.; Anderson, C.M.; Polcari, A.; Glod, C.A.; Maas, L.C. & Renshaw, P.F. (2000). Functional deficits in basal ganglia of children with attention-deficit/hyperactivity disorder shown with functional magnetic resonance imaging relaxometry. *Nature Medicine*, *6*, 470–473.

Tobler, P.; Fiorillo, C. & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642-5.

Triesch, J.; Jasso, H. & Deák, G. (2007). Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior*, *15*(2), 149–65.

Ungless, M.; Magill, P. & Bolam, J. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, *303*, 2040-2.

Vessel, E. (2004). *Behavioral and neuralinvestigation of perceptual affect*. Unpublished doctoral dissertation, University of Southern California.

Weber, C.; Muse, D.; Elshaw, M. & Wermter, S. (2005). Reinforcement learning in MirrorBot. In *Proceedings of 15th international conference on artificial neural networks* (p. 305-10). Springer-Verlag Berlin Heidelberg.

Weber, C. & Triesch, J. (2006). A possible representation of reward in the learning of saccades. In *Proceedings of the 6th international workshop on epigenetic robotics* (pp. 153–60). Lund University Cognitive Studies.

Weber, C.; Wermter, S. & Elshaw, M. (2006). A hybrid generative and predictive model of the motor cortex. *Neural Networks*, *19*(4), 339–53.

Weber, C.; Wermter, S. & Zochios, A. (2004). Robot docking with neural vision and reinforcement. *Knowledge-Based Systems*, *17*(2-4), 165-72.

Weigmann, K. (2006). Robots emulating children. *EMBO Report*, *7*(5), 474-6.

Wermter, S.; Weber, C. & Elshaw, M. (2004). Associative neural models for biomimetic multi-modal learning in a mirror neuron-based robot. In A. Cangelosi, G. Bugmann & R. Borisyuk (Eds.), *Modeling language, cognition and action* (p. 31-46).

Wermter, S.; Weber, C.; Elshaw, M.; Gallese, V. & Pulvermüller, F. (2005). Biomimetic neural learning for intelligent robots. In S. Wermter, G. Palm & M. Elshaw (Eds.), (p. 162-81). Springer.

Wermter, S.; Weber, C.; Elshaw, M.; Panchev, C.; Erwin, H. & Pulvermüller, F. (2004). Towards multimodal neural robot learning. *Robotics and Autonomous Systems*, *47*(2-3), 171-5.

Wheeler, M. (2005). *Reconstructing the world*. MIT Press.

Wilson, C. J. (2004). Basal ganglia. In *The synaptic organization of the brain* (pp. 361–415). Oxford University Press.

Wood, R.; Humphries, M. & Gurney, K. (2006). A large scale biologically realistic model of the neostriatum. In *Computational Neuroscience meeting (CNS).*

Reinforcement Learning

Edited by Cornelius Weber, Mark Elshaw and Norbert Michael Mayer

ISBN 978-3-902613-14-1
Hard cover, 424 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2008
Published in print edition January, 2008

Brains rule the world, and brain-like computation is increasingly used in computers and electronic devices. Brain-like computation is about processing and interpreting data or directly putting forward and performing actions. Learning is a very important aspect. This book is on reinforcement learning which involves performing actions to achieve a goal. The first 11 chapters of this book describe and extend the scope of reinforcement learning. The remaining 11 chapters show that there is already wide usage in numerous fields. Reinforcement learning can tackle control tasks that are too complex for traditional, hand-designed, non-learning controllers. As learning computers can deal with technical complexities, the tasks of human operators remain to specify goals on increasingly higher levels. This book shows that reinforcement learning is a very dynamic area in terms of theory and applications and it shall stimulate and encourage new research in this field.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821