Università degli Studi di Salerno

Dipartimento di Ingegneria Elettronica ed Ingegneria Informatica

Dottorato di Ricerca in Ingegneria dell'Informazione
XV Ciclo – Nuova Serie

TESI DI DOTTORATO

A neurocomputational model of reward-based motor learning

CANDIDATO: **GIANMARCO RAGOGNETTI**

TUTOR: **PROF. ANGELO MARCELLI**

CO-TUTOR: **PROF. ANDREA VIGGIANO**

COORDINATORE: **PROF. ALFREDO DE SANCTIS**

Anno Accademico 2016 – 2017

Acknowledgements

The more I see, the more I know,

the more I know, the less I understand.

Quote from a comic book which made my childhood

Doing a proper thanks to everybody accompanied me during this three-year journey could be a tough task, way more than writing a thesis, yet I am not so skilled in saying 'thank you' to people, so I will try to make my best effort without going too much on the mawkish side.
First, I want to thank my mother, Carmen who has been both in good and bad times a great support and always believed in my abilities.
I would thank my beloved grandma, Lucrezia, my second mom and my first teacher, which introduced me to the discipline of study and always took care of my learning route.
I would thank my friends, which provided the right balance of confrontation, interior enrichment (lot critiques and mindless distractions.
I would thank my neighbors who behaved like an extended family and had the patience to volunteer for hours of experiments, pivotal for the results of this work.
The same for all the students from Medicine Faculty , Engineering Faculty and other volunteers a bit above the age to be university students…
I left for last people, which were directly involved in this work.
Professor Viggiano, which proved a priceless mentor and gave me basis of physiology, robotics and bricolage (quite eclectic for a physician).

Letizia and other collaborators from Medicine Faculty which helped me gathering the data for the experiments
I want to thank people in my lab, which provided a peaceful and stimulating atmosphere and helped me debugging most of the software.
Danilo, for showing a lot of patience to explain me how to port software from Emergent to Matlab.
Adolfo, for helping me remembering notions of electronics while building Arduino experiments, and always calling for a coffee break in the right moments;
Rosa, for helping me understanding the rules of Emergent and how neural networks actually work, and for bringing cakes and a tea boiler to ease our tough times of brainstorming;
Canio, for helping me understanding programming and keeping updated with latest tech trends;
Antonio, for being always on place to save the day both as a precious colleague and a friend.
I want to thank all the trainees, which alternated in our lab and provided their enjoyable presence and knowledge and the cleaning ladies for giving us the excuse to make a break.
Above all Professor Marcelli, which teach me how someone curious about science and life should behave, to adopt an active and open minded attitude toward all fields,for spoiling us with exquisite homemade recipes and for providing always a paternal and benevolent enlightenment aura.
To quote Looney Tunes, ‘That’s’ all Folks’!

Contents

Introduction

Computational models of brain motor learning provide useful tools for validating the hypothesis and showing how simulations of learning could work. Previous works dealt with the neural structures involved in learning a new motor skill and reaching movements. Aim of this PhD thesis is to add a further piece to this complex framework.

Learning to perform a movement is a complex action that involves different areas of central nervous system.

According to behaviorism, learning a new movement, voluntary or not, is the ability to change the probability of a specific motor response, given one or more sensory stimuli.

Differently from *reflexes*, which are innate, involuntary motor responses, learning a voluntary movement is driven by performance evaluation, which helps to produce increasingly accurate movements.

In this thesis, we address the following main questions:

- Where does this evaluation mechanism originate?
- How does this value system work?
- Which are the main brain areas involved?
- How quickly does a human learn a skill trough?

As concerns the first question, we will show that the performance evaluation, which could be for better or worse, helps narrowing down the field of all possible movements so that only specific ones will be able to fulfill the assigned task.

We call this evaluating stimulus *reward* in the former case, *punishment* in the latter. Some types of rewards/punishments always reinforce/weaken learning. They arise from innate drives, such as hunger, thirst, physical pain, sexual arousal, etc., originate from specific sites and come somehow hardwired in our brain. We call them *primary values*. Yet, reinforcing/punishing stimuli could also originate from concepts that are more sophisticated, like money, fame and overall approval. These concepts at first have a neutral meaning, but they can acquire the power to drive actions trough learning. For this reason, we can call them *learned values*. How this learning of new values could happen? We believe that an initial neutral stimulus could become a learned value after repetitive pairings with a primary value in a classical conditioning paradigm.

Regarding the second question, we could assume that mechanism of learning occurs in a trial and error fashion: in the first stages when the movement has still to be learned, all the possible actions/movements have the same probability, but as the learning process goes on, the reinforcement/punishment alters the probability of some actions in favors of others. A reinforcement /punishment could be delivered as a direct consequence of the action performed (*operant or instrumental conditioning*), but could also be completely uncorrelated with its semantics, in an associative paradigm (*classical conditioning*): for example, an infant could wave his hands or cry in order to get attentions from the parents.

As for third question, a common belief is that there are two main architectures in the brain responsible for the role of reinforcement learning:

- a system assigned for storing and selecting actions, made up of Cerebral Cortex, mapping all possible combinations of sensory inputs with all possible motor actions, and Basal Ganglia which acts as a gating device to enable/block specific actions;
- a system assigned to produce evaluations, which occurs in form of dopamine firing to the Basal Ganglia, and is able to modulate the gating action of the previous system.

Primary values originates from special sensory input (e.g. *lateral hypothalamus* for reward and *lateral habenula* for punishment) and always produce a *dopamine* emission, while learned values could originate from sensory inputs or from Cortex, and learn to produce dopamine after repetitive pairing with primary values, in a Pavlovian conditioning paradigm.

The neural structures of dopaminergic system are supposed to be the Amygdala for the learning of values and Ventral Tegmental Area (VTA) for the production of dopamine.

The learning rule requires establishing the amount of changes in the synaptic weights. This consists often of a parameter, *learning rate,* chosen in such a way to obtain a good performance of the model.

As regards the last question, we performed several experiments in order to evaluate the speed of human learning, providing realistic values for the learning rate coefficient.

In the thesis, we will investigate all those aspects and propose two main contributions:

- a model of the brain areas involved in learning by reward;
- a computational model to validate the model.

Accordingly, the thesis unfolds as follows:

- *Chapter 1* explores the body of knowledge and the state of the art concerning learning by reward and provides the foundations and basic assumptions for our work.
- *Chapter 2* illustrates the modeling of learning by rewards and provides physiological and anatomical descriptions of the main neural structures involved.
- *Chapter 3* introduces the computational model developed for validating the working hypothesis the model of the previous chapter is based upon.
- *Chapter 4* shows the experiments performed in order to evaluate realistic human learning rates, and the achieved results.
- Eventually, the *Conclusions* summarize the main findings and outline possible future development to incorporate into the model other brain areas

Chapter 1

Functional aspects and neurobiology of reward based behavior and learning

Behavior is in general an intuitive concept; still it is difficult to find an exact and universal definition of behavior. It is generally acknowledged that an (animal) behavior consists of a sequence of movements that is influenced by an ensemble of sensory inputs [1]. In *applied behavioral analysis* (ABA), the ensemble of sensory inputs is called *stimulus* and the sequence of movements, which in turn are sequences of motor activations, are called *responses* [2].

Therefore, a movement in its most elementary form is a *stimulus- response association.*

We will review some aspects of learning supporting the claim that hierarchical loops exists in the brain and the most generic high level paradigm is a *four term contingency*[2]. Within this paradigm, learning could occur at different stages.

1.1. Learning of movements in human brain: from reflexes to voluntary actions

In the brain, movements are organized in a hierarchical fashion starting with *innate reflexes* automatically triggered in order to fulfill concrete needs, i.e. nutrients required to maintain the body, to the abstract *goals directed behaviors* like having dinner in a specific restaurant or learning a new language.

At the lowest level , there are movements which immediately follow a particular sensory input (within 0.1 sec) and generally consist of simple, *non-voluntary* actions (stimulus-response pairing); this kind of movements are called *reflexes* and their neural substrate is well known [3]. These stimuli could come from the environment or from internal state of body, defining *needs*.

According to ABA terms, these somewhat fast, non-voluntary, stimuli-evoked movements are called *respondent behavior* [2]. It can be acknowledged that there are many innate (primitive) reflexes, which are structurally present at

birth, such as, for instance, the *myotatic reflex*, the *righting reflex*, the *withdrawal reflex* or the *palmar grasp reflex* [4].

It can also be acknowledged that some new reflexes can be learned. For example, an acoustic tone can produce an *eye-blink reflex*, if previously paired many times with an eye puff, which produces a primitive eye-blink reflex (Christian and Thompson, 2003). This kind of new generated reflexes, which is a form of learning, are generally called *conditioned reflexes*, and the learning process is called *classic conditioning* [6]. According to ABA terms, this kind of learning is called respondent conditioning [2].

Further ascending the hierarchy, movements consist of a more sophisticated sequence of actions which do not immediately follow a particular sensory input and can neither be predicted for sure by a particular ensemble of sensory inputs (stimulus); this kind of movements are commonly called *voluntary movements*. Despite an exact prediction cannot be made, it can be admitted that, given a particular stimulus, the frequency of some voluntary movements are greater than others and that such possibility can also change over time. According to ABA terminology, this type of movements are called *operant behaviors*.

Observing the learning process of voluntary movements, we can be recognize particular types of stimuli that cause the change in the frequency of the movements being executed; these particular types of stimuli are generally called *reinforcers* (usually in behavioral analyses) or *rewards* (usually from a psychological point of view) and punishments. A reinforcer is a particular stimulus that will increase the frequency of an ongoing operant behavior, i.e. a particular voluntary response associated to a particular

stimulus; conversely, a punishment will decrease the probability of a particular operant behavior [2].

A reinforcer selects not just certain forms of behavior; it also selects the environmental conditions that in the future will evoke instances of the response class. A behavior which occurs more frequently under some antecedent conditions than it does under others is called *discriminant operant*. The specific stimulus that triggers the discriminant operant behavior is called *discriminative stimulus* (S^D). Therefore, the *three-term-contingency* (antecedent, behavior and consequence) is the basic unit of operant behavior, as shown in figure 1.1

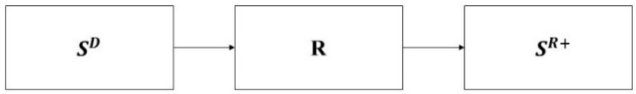


Figure1.1 In a 3 term contingency paradigm, reinforcer stimuli S^{R+} strengthens the association between a response R and a specific discriminant stimulus S^D $\mathrm{S^{R+}}$.

The way a reinforcer/punisher is delivered could be the consequence of that specific action (i.e. pressing a lever in order to obtain food) or it could be uncorrelated from the behavior performed, (i.e. a child could receive a prize if he fulfilled a task assigned by parents, like keeping his room clean or doing all his homework). The former case is still a classical conditioning scenario; the latter is another type of learning, called operant (instrumental) conditioning [2, 7].

As there is a hierarchy of movements, so there is a hierarchy of stimuli. The term *unconditioned stimulus* refers to a stimulus that always elicits a reflex or in the case of

voluntary movements always acts like a reinforcer, reward or punishment. The term *conditioned stimulus* usually refers to an initially neutral stimulus that did not produce a reflex until it has been paired several times with an unconditioned stimulus. In learning voluntary movements, a conditioned stimulus could acquire the ability to reinforce/punish a behavior after being paired with an unconditioned rewarding/punishing stimulus. Eventually, a conditioned stimulus in turn can act as an unconditioned stimulus to a new stimulus in what is called *second order conditioning*. Figure 1 2 depicts the relations between the chance in the stimulus and the effect of the behavior performed in defining the role of a stimulus as a reinforcer or a punisher, either positive or negative.

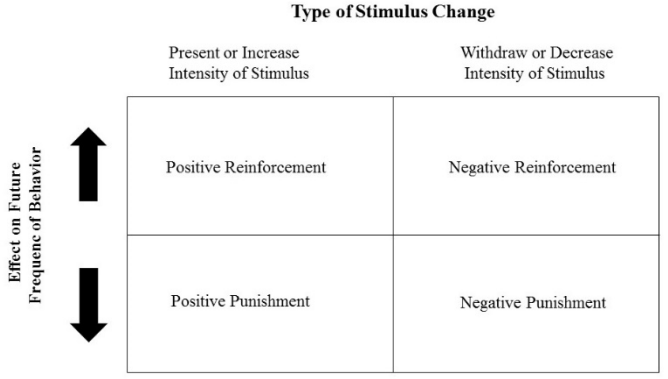


Figure 1.2. Positive and negative reinforcers and punishers are defined by the type of stimulus change operation that immediately follows a behavior and the effect that operation has on the future frequency of that type of behavior

1.2 Learning of reinforcers and punishers: primary and learned values

As mentioned in the previous section we could make a distinction among rewards/punishers. In other words, there are some innate rewards and punishment arising from primary needs of the body (e.g. feeding and pain), but other stimuli, that have no reinforce/punishment effect at first (e.g. money, or scolding), can acquire this kind of effect after being paired with an innate reinforcer/punishment.

We can refer to the first group as to primary values, while to the second group as learned values (or conditioned reinforcement/punishment). Usually conditioned reinforcers occur naturally in the learning of a new movement, in a second order conditioning fashion: a voluntary movement could be seen as a sequence of motor responses (actions) each occurring under certain stimuli conditions states. Innate reward (e.g. food) is provided at the end of the sequence, so after repetitive training, intermediate sensory states (discriminative stimuli, see previous section) become reinforcers for the actions which lead in those states.

For example, the sight of the ice cream truck or of the indication to a restaurant becomes a conditioned reinforcer for food (innate reward); similarly in a punishment scenario, the sight of an advertisement ‘Do Not Touch, High Voltage’ or

hearing the noise of a wild animal could incentivize the act of avoiding unpleasant encounters.

1.3 The role of context and motivating operations

The existing level of *motivation* can influence the efficacy of reinforcers/punishers and the likelihood of a conditioned response. Motivating operations (MOs) are environmental variables that have two effects on behavior [8]:

- altering the *operant reinforcing effectiveness* of some specific stimuli (value-altering effect);
- altering the *momentary frequency* of all behaviors that have been reinforced by those stimuli (behavior-altering effect)

Motivating operations can take two forms: *establishing operation* (EO), when the MO increases the effectiveness of a reinforce (e.g. food deprivation makes food more effective as a reinforcer), or *abolishing operati*on (AO) when the MO decreases the effectiveness of a reinforce is an (e.g. food ingestion reduces the effectiveness of food as a reinforcer).

Adding the MO and namely its EO form to the *three-stage contingency* of paragraph 1.1 results in the *four-term contingency* shown in figure 1.3.

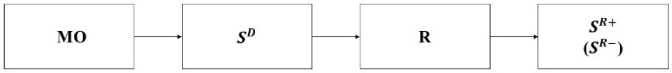


Figure1.3. A motivating operation (MO) can increase/decrease the effectiveness of a stimulus as reward.

Prior to discussing MOs in detail, it is important to draw a clear distinction between MOs and another class of antecedents, *discriminative stimuli*. Discriminative stimuli are events that have been associated with the availability or non-availability of reinforcement in the past. The presence of a green light on a drinks dispenser, for example, signals the availability of a can of soda, whereas the presence of a red light signals its unavailability.

Going back to MOs, there are some innate *unconditioned motivating operations* (UMO).Cooper [2] finds out there are nine *UMOs*: deprivation of food, water, sleep, activity, sex or oxygen, becoming too warm or too cold, and increase of a painful stimulation.

There are also **conditioned** motivating operations (CMO) that result from the learning history of the organism. Three kinds of conditioned operations [2] have been identified: *surrogate*, *reflexive*, and *transitive*.

A *surrogate CMO (CMO-S)* accomplishes the same value altering and behavior altering effects as the MO it was paired with when it was learned. Consider, for example, a person who always has lunch at midday. The time on the clock in addition to having discriminative properties (such as signaling the opening of the canteen) may also exert a motivating influence. Following the repeated pairing of food deprivation and the time of 12:00 p.m. on a clock, the time on the clock may eventually acquire motivating properties of its own. That is, through repeated association with an UMO (food deprivation acting as a EUO), the previously neutral stimulus (time on the clock) may itself establish the reinforcing value of food and evoke food-related behavior independent of actual levels of food deprivation. The time on the clock may also establish the punishing value of food unavailability and reduce behaviors that have been associated with such delays in the past, such as answering the telephone, independently of current levels of food deprivation.

A *reflexive CMO (CMO-R)* acts as a reinforcement when it is removed. The onset of a *CMO-R* is associated with either the *improvement* or *worsening* of the person's condition. Therefore, its onset alters the value of its own removal (or continued presence) as a type of reinforcement (or punishment) and alters the probability of behaviors occurring that have previously been associated with these consequences. The CMO-R therefore acts on its own reinforcing value and not on that of another stimulus (as is the case with the CMO-S). Take a young infant for whom the onset of certain social stimuli (such as seeing his or her mother frown) is correlated with the subsequent onset of an aversive stimulus, such as being scolded and thus the ‘worsening’ of his or her condition.

The onset of the mother's frown may establish its own offset as an effective form of reinforcement and evoke behaviors that have been associated with its removal in the past, such as the infant beginning to cry or ceasing the activity in which he or she was engaged, thereby acting as a reflexive conditioned establishing operation (CEO-R).

A *transitive CMO (CMO-T)* makes something else effective as reinforcement. An example of a transitive conditioned establishing operation *(CEO-T)*, typically seen in approaches such as incidental teaching [10], involves contriving a situation in which one stimulus increases the value of a second stimulus as a type of reinforcement. The second stimulus cannot be obtained until a given behavior has occurred [9]. A *CMO-T* relation may be present when an ongoing response or behavior chain (such as purchasing a soft drink) is blocked or interrupted (perhaps by having the incorrect change). In such circumstances, the initial stimulus change, which would normally function as a discriminative stimulus for the now blocked response (such as the sight of the drink dispenser), instead functions as a *CMO-T*. It establishes the reinforcing value of a second stimulus change (such as getting correct change for the machine). This *CMO-T* evokes a second response that has been effective in achieving this second stimulus change in the past (such as asking the shop assistant for some change). The initial stimulus change (sight of the drinks machine) acts as a *CEO-T* for the second stimulus change (getting the correct change) and alters behavior accordingly. The *CMO-T* is conditional and would only be expected to exert any influence when an *EO* is in effect for the terminal response (such as when the person is ‘thirsty’).

1.4 Putting the things all together: goal directed behaviors and the H4W problem

Up to this point, we explored how a voluntary movement occurs and which are the learning paradigms. Now we want to move to the top of hierarchy . It must be assessed voluntary actions can be either habitual or goal-directed. In order to be considered goal-directed, a behavior must satisfy two requirements:

- the individual should display knowledge of the causal efficacy of its own actions and their outcomes (reinforcement/punishments) given the current state (discriminant stimuli) or context (motivating operation);
- the individual should select and regulate its behavior using goal representations, e.g. internally generated representations of desired action outcomes.

Goal directed behavior is distinct from other kind of control, such innate reflexes and habits, in the sense that it does not describe a specific operation or procedure but rather the end state an operation should achieve. A goal directed behavior depends on tightly coupled processes that involve perception, motivation, emotion, cognition and action. It cannot be localized to a 'central goal nucleus' in the brain, but rather depends on the interplay of a number of mechanisms realized

in several brain areas. According to [11], in order to act in the external world the brain needs to set of objectives that are captured in answering the questions:

- *Why* do I need to act?
- *What* do I need?
- *Where* and *When* can this be obtained?
- *How* do I get it?

These five questions can be defined as the *H4W problem* [11]. In short, animal (human) needs to determine a behavioral procedure to achieve a goal state *(How)*, which in turns requires defining the motivation in terms of needs, drives, and goals *(Why)*. It also requires information on the objects and their affordances in the world *(What)*, on the location of the objects and self in the world, i.i the spatial configuration of the task domain *(Where)*; eventually the sequencing and timing of action relative to the dynamics of the world and self *(When)*.

In the next chapter, we will find out how these questions are answered in our brains.

1.5 Final behavior considerations

Most of the times, experiments with classic conditioning are aimed to observe new generated reflexes, while operant conditioning is aimed to explore changes in voluntary movements. Nevertheless, it should be remembered that the three kind of learning (new reflexes, voluntary movements, rewards) are not mutually exclusive and it is very difficult to imagine a situation in which only one of these three processes takes place, unless the neural circuit responsible for one of

them is damaged. This is exactly how the observation of the consequence of local lesions can give important information about the role played by each part of the nervous system in the aspects of movement control and learning depicted above. The information obtained in this way will be revised in the next chapter.

The last behavioral consideration regards the global effect that emerges from the parallel development the three kind of learning above postulated Looking at the movements of a child at birth, it can be recognized that there are lot of reflexes already working; these can be called primary reflexes [4].

Beside these reflexes, the voluntary movements appear to consist mainly of random movements. For example, the suction reflex is already present, but pointing the head toward the nipple is mainly a random event; the child would never be able to reach the nipple in the first day of his life, if the mother does not put her nipple near the child's mouth. When the nipple touches the child's cheek on one side, the first times voluntary movements will consist of randomly turning the head toward one side or the other. Whenever the right movement occurs and the nipple enters the child's mouth, the suction reflex starts; the milk will enter the mouth, the swelling reflex also occurs. At this time a reward signal, due to feeding, will increase the chance of voluntarily activating that head movement when a touch sensation is present on that cheek side and when a hungry sensation is present. Repeating this situation, the chance of that voluntary movement will increase more and more. Thus, it will happen that that touch sensation will be associated many times with that particular head movement; this repeated association will produce a new automatic, non-voluntary reflex response, so, at a certain

point, the cheek touch on one side will produce a fast, non-voluntary turning of the head to that side. In summary, rewards will guide voluntary choices of movements in each situation (i.e. given a particular stimulus); such voluntary responses have a certain amount of delay due to the intense elaboration of incoming stimuli. At the same time, whenever a particular couple of stimulus-movement is repeated many times, such association is retained and executed faster by a non-voluntary reflex mechanism; by this way, many non-voluntary conditioned reflexes will be acquired and will result in "accurate" movements because they follow the stimulus with a very small delay.

1.6 Reinforcement learning rules

In the following sections, a review of the main algorithms for reinforcement learning from literature will be provided. We will show how reinforcement-learning algorithms lie in the framework of learning methods.

In Reinforcement Learning (RL) an agent learns from the consequences of its actions, rather than from being taught (supervised learning), and selects its action on basis of its experience. The reinforcement signal the agent receives is a numerical reward, so the agent will learn to increase that action, which led to reward, or maximize reward in the case reward is cumulative. According to Marr's approach (Marr et al 1976 later re-introduced by Gurney [13]), there are three main different levels: an algorithmic more abstract approach used in machine learning field, a mechanistic more directed toward neural networks, and an implementation level, which is biologically grounded .Figure 1.4, depicts the framework.

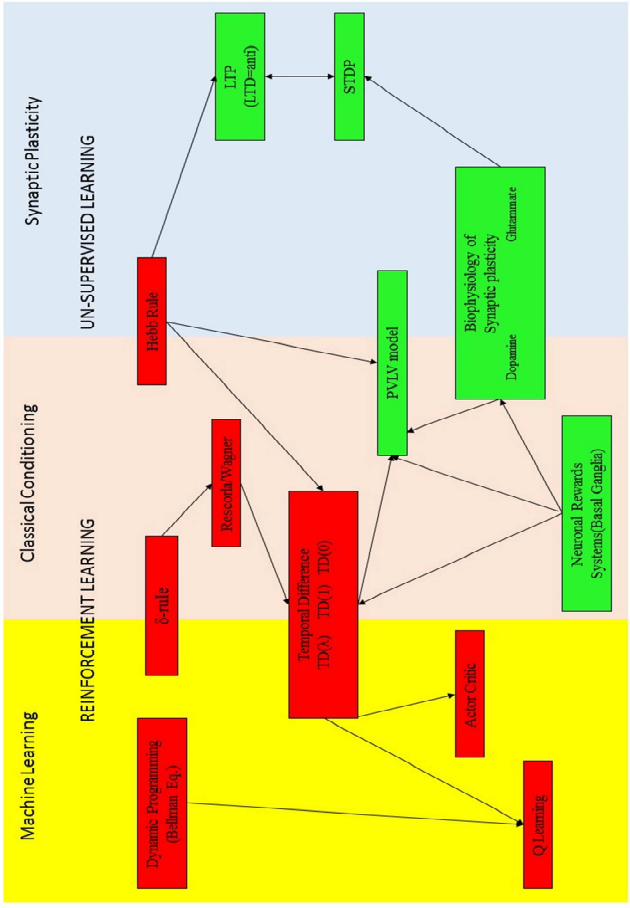


Figure 1.4. A diagram of the framework of reinforcement learning depicting the links between the different fields. Red blocks refers to the most important theoretical models, green boxes to the biologically inspired ones.

1.6.1 The algorithmic level :the Machine Learning perspective

Reinforcement Learning can be formulated as class of *Markov Decision Problems (*MDP). The agent can visit a finite number of *states* and in visiting a state, a numerical *reward* will be collected, where negative numbers may represent punishments. Each state has a changeable *value* attached to it. From every state, there are subsequent states that can be reached by means of *actions*. The value of a given state is defined by the *averaged future reward*, which can be accumulated by selecting actions from this particular state. Actions are selected according to a *policy*, which can also change. The goal of an RL algorithm is to select actions that maximize the expected cumulative reward (the *return*) of the agent.

RL methods are employed to address two related problems: the *Prediction Problem* and the *Control Problem.*

- *Prediction only:* RL is used to learn the *value function* for the policy followed. At the end of learning this value function describes for every visited state how much future reward we can expect when performing actions starting at this state.
- *Control:* By interacting with the environment, we wish to find a policy, which maximizes the reward when traveling through state space. At the end, we have obtained an *optimal policy*, which allows for action planning and optimal control. Since this is really a

predictive type of control, solving the control problem would seem to require a solution to the prediction problem as well.

In general there exist several ways for determining the optimal value function and/or the optimal policy.

If we know the state *transition function T(s, a, s')*, which describes the transition probability in going from states to s' when performing action a, and if we know the *reward function r(s, a),* which determines how much reward is obtained at a state, then algorithms are called *model based algorithms*. They can be used to acquire the optimal value function and/or the optimal policy. Most notably *Value-Iteration* and *Policy-Iteration* are being used, both of which have their origins in the field of *dynamic programming* [14] and are, strictly-speaking, therefore not RL algorithms [21].

If the model of the process is not known in advance, i.e. both the transition and the reward functions are unknown, and then we are truly in the domain of RL, where by an adaptive process the optimal value function and/or the optimal policy will have to be learned. Machine Learning and instrumental conditioning deal with closed-loop control problems.

However, Classical Conditioning deals with a prediction-only problem because the response of the animal does not influence the experiment or, in more general terms, does not influence the environment. A good short summary relating algorithmic approaches to real classical conditioning experiments is given in [15].

Arising from the interdisciplinary study of these two fields, there appeared a very influential computational method, called the method of *Temporal Difference Learning*

(TD) [24, 31]. TD learning was originally mainly associated to animal learning (Classical Conditioning). It was essentially the work of Klopf [16, 17, 18, 19] that began to bring TD-methods together with animal learning theories

1.6.2 The mechanistic level :the Neuronal perspective

The state-action space formalism used in reinforcement learning can be translated into an equivalent *neuronal network* formalism, as will be discussed below.

Preliminarily, let us note that the neuronal perspective of RL is indeed meant to address biological questions. Its goals are usually not related to those of other *artificial neural network* (ANN) approaches, which are addressed within the machine-learning approach.

1.6.2.1 Rescorla Wagner rule

One of the simplest learning paradigm comes from the application of delta rule [22] to reinforcement learning and is known as the Rescorla Wagner model *(R-W)*. It is a model of classical conditioning: no knowledge of the actions is necessary, and learning is conceptualized in terms of associations between conditioned (CS) and unconditioned (US) stimuli. A strong CS-US association means, essentially, that the CS anticipates the US. Thus, the delta rules becomes proportional to the difference between the reward and the

expected value of reward, which is calculated on the bases of stimuli activations and their synaptic weights.

Figure 1.5. According to RL, the weight of synapse between an input stimulus S and a reinforcement stimulus r is strengthened as the learning process progresses. Rescorla Wagner learning rule defines that the rate of this growth is proportional to the difference between reward and a prediction made by the stimulus (or a sum of stimuli in the case of multiple inputs).

The learning rule for updating the synaptic weights is:

$$w_i = w_i + \in \delta S_i \ ;$$

where $\in$ is the learning rate and $\delta = r - \sum_i w_i S_i$.

One of the biggest issues with the R-W model is that it does not model learned reward; in such scenarios, a second order conditioning occurs between a new stimulus and the CS, which should act as a reward. This problem, which goes under the name of *temporal credit assignment*, needs a representation of time, which will be resolved by the algorithm described in the next section**.**

1.6.2.2 Temporal Difference learning rule

The Temporal Difference model (Sutton) incorporates a prediction of future rewards. It does so by adding to the Rescorla-Wagner model one additional term to the delta equation, representing the future reward values that might come later in time:

$\delta = r - \sum_i w_i S_i + f$

where f represents the future rewards. Reward expectation, now, has to try to anticipate both the current reward r and the future reward f. In a simple conditioning task, where the CS reliably predicts a subsequent reward, the onset of the CS results in an increase in this *f* value, because once the CS arrives, there is a high probability of reward in the near future.

Furthermore, this *f* itself is not predictable, because the onset of the CS is not predicted by any earlier cue (and if it were, then that earlier cue would be the real CS, and drives the dopamine burst). Therefore, the reward expectation cannot cancel out the *f* value, and a dopamine burst ensues.

Although this *f* value explains CS-onset dopamine firing, it raises the question of how can the system know what kind of rewards are coming in the future? Like anything having to do with the future, it fundamentally is just a guess, using the past as a guide as best as possible. TD does this by trying to *enforce consistency in reward estimates over time*. In effect, the estimate at time *t* is used to train the estimate at time *t-1*, and so on, to keep everything as consistent as possible across time, and consistent with the actual rewards that are received over time.

This can all be derived in a very satisfying way by specifying something known as a *value function V (t)*, which assigns values to states and then calculates the change of those values by means of a temporal derivative. As a consequence, these methods are related to correlation-based, differential Hebbian learning methods, where a synaptic weight changes by the correlation between its input signals with the derivative of its output.

V (t) is the sum of all present and future rewards, with the future rewards *discounted* by a *"gamma"* factor, which captures the intuitive notion that rewards further in the future are worth less than those do that will occur sooner:

$$V(t)=r\,(t) +\gamma^{1} r\,(t+1) + \gamma^{2} r\,(t+2)$$

We can get rewrite the equation in a recursive way:

$$V(t)=r\,(t) +\gamma\, V\,(t+1)$$

So we can consider

$$f= \gamma\, V\,(t+1)$$

Figure 1 6 shows how the prediction of rewards occurs in TD learning; the production of *V (t)* needs time as a representation of input that is time should be coded as a sequential activity of the set input neurons.

Sutton has also proposed an updated version of TD learning named TD Lambda, denoted with *TD (λ)*. Here the presence of eligibility traces is taken in account. Eligibility traces model how a CS input stimulus does not abruptly go to baseline, but instead slowly decays, so for a period CS and US (reward) occur simultaneously. Eligibility traces are usually implemented by an exponentially decaying memory trace,

with decay parameter λ. This generates a family of TD algorithms TD *(λ), $0 \leq \lambda \leq 1$*, with TD (0) corresponding to updating only the immediately preceding prediction as described above, and TD *(1)* corresponding to equally updating all the preceding predictions.

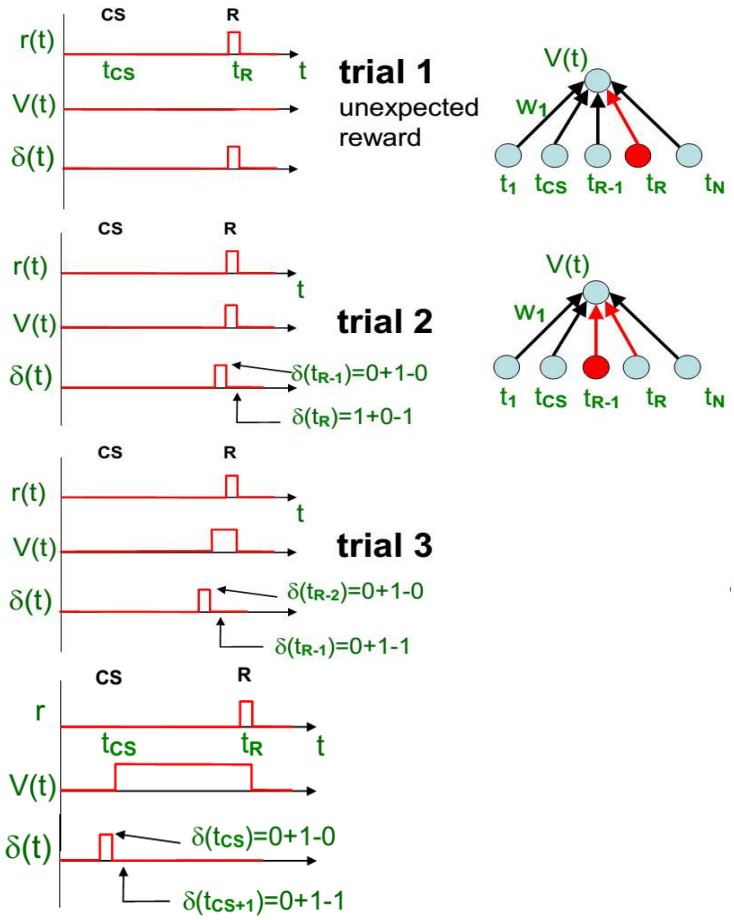


Figure 1.6. In Temporal Difference Learning, a prediction of reward r occurs over time in form of a value function V (t), which is updated on each trial and this pulls the delta signal back in time. The input to V (t) is a representation of time.

1.6.2.3 Actor Critic Architecture

The algorithms explored thus far cover a prediction problem, now we want to close the loop by exploring how prediction can drive the choice of an action.

Neuronal approaches, which address the control problem and can generate behavior, mostly follow a control-loop architecture. Figure 1.7 shows a conventional feedback control system. In neuronal terms, this is a reflex-loop. A controller provides control signals to the system, which is influenced by disturbances. Feedback allows the controller to adjust it signals.

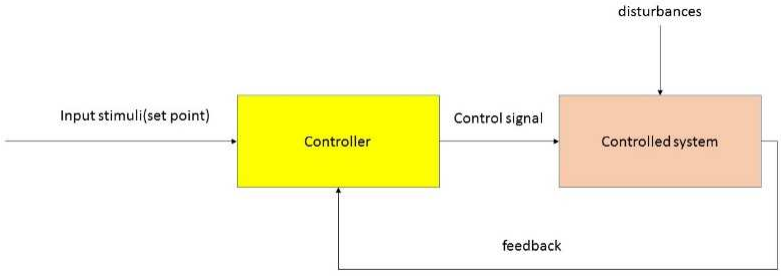


Figure 1.7. A controlled feedback loop (reflex). The input sets the position and the controller learns, through information provided by feedback, to bring the controlled system to the desired position.

This could be extended into an Actor-Critic architecture [24, 25] The Critic produces evaluative, reinforcement feedback for the Actor by observing the consequences of its actions. The Critic takes the form of a TD-error δ, which gives

an indication if things have gone better or worse than expected with the preceding action. If the TD-error is positive the tendency to select this action should be strengthened, otherwise weakened. Thus, Actor and Critic are adaptive through reinforcement learning.

On the machine learning side, Actor-Critic is related to interleaved value/policy-iteration methods [21]. On the side of control, they are related to advanced feed-forward control and feed-forward compensation techniques. Assumptions were made that the input states can produce both the prediction, instead of time, and at the actions so they are inputs to the Critic and the Actor blocks. Figure 1.8 depicts the actor-critic model from both the Machine Learning and the Neuronal perspective

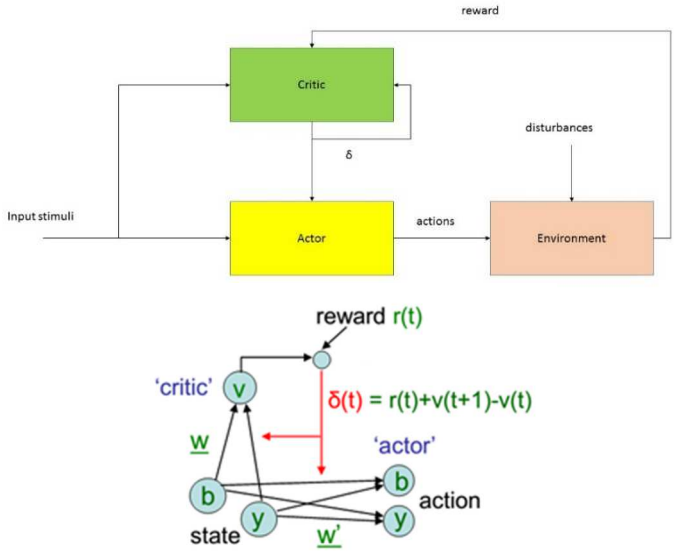


Figure 1.8. The Actor- Critic model makes use of the δ-TD reinforcement signal coming from the Critic in order to optimize a policy of actions in the Actor

The Actor Critic architecture solves *the structural credit assignment problem*, which consists in trying to maximize the expected return by choosing the best actions. The Actor uses in general a set of predefined actions. Actions are not easily generated *de novo*. The Critic cannot generate actions on its own but must work together with the Actor. Convergence is slow if these methods are not augmented by additional mechanisms .

1.6.2.4 Q-Learning

Q learning [26] is a machine learning algorithm that can be used to find an optimal action-selection policy for any given (finite) *Markov decision process (MDP)*. It works by learning an action-value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter. A policy is a rule that the agent follows in selecting actions, given the state in which it is located. When such an action-value function is learned, the optimal policy can be constructed by simply selecting the action with the highest value in each state. One of the strengths of *Q*-learning is that it is able to compare the expected utility of the available actions without requiring a model of the environment.

Additionally, *Q*-learning can handle problems with stochastic transitions and rewards, without requiring any adaptations. It has been proven that for any finite *MDP*, *Q*-learning eventually finds an optimal policy, in the sense that the expected value of the total reward returned over all the successive steps, starting from the current state, is the maximum achievable.

The algorithm takes in account an agent, states S and a set of actions A per state. By performing an action $a \in A$, the agent can move from state to state. Executing an action in a specific state provides the agent with a reward (a numerical score). The goal of the agent is to maximize its total reward.

It does this by learning which action is optimal for each state. The action that is optimal for each state is the action that

has the highest long-term reward. This reward is a weighted sum of the expected values of the rewards of all future steps starting from the current state, where the weight for a step from a state Δt steps t into the future is calculated as $\gamma^t \gamma \Delta t$. Here again γ is a discount factor ($0<\gamma<1$) and trades off the importance of sooner versus later rewards. γ may also be interpreted as the likelihood to succeed (or survive) at every step Δt .

The algorithm therefore has a function that calculates the Quantity of a state-action combination. Before learning has started, Q returns an (arbitrary) fixed value, chosen by the designer. Then, each time the agent selects an action, and observes a reward and a new state that may depend on both the previous state and the selected action, Qis updated:

$$Q(s_t , a_t) = Q(s_t , a_t) + \alpha (r_{t+1} + \gamma Q(s_{t+1} , a_{t+1}) - Q(s_t , a_t))$$

where α is the *learning rate* and r_{t+1} is the reward observed after perfoming a_t in s_t . *Q*-learning looks similar to TD formalism, the difference is that we are visiting state *s* from where we take the specific action *a*, whereas in TD the action was left unspecified.

For the Q- learning algorithm no neuronal architecture has been proposed so far.

1.6.3 The implementation level :the Neuroscience perspective

In this section, it will be established the link between the mechanistic level and the neuroscience, hence establishing a link between the abstracts ANNs presented in the previous sections with neurophysiological findings.

In general the Dopaminergic system of the brain is held responsible for RL. Responses from dopaminergic neurons have been recorded in the Substantia Nigra pars compacta (SNc) and the Ventral Tegmental Area (VTA) where some reflect the prediction error δ of TD-learning [27]. These neurons have been discovered mostly in conjunction with appetitive (food-related) rewards. Figure 1.9 shows some examples of prediction error- as well as reward expectation neurons.

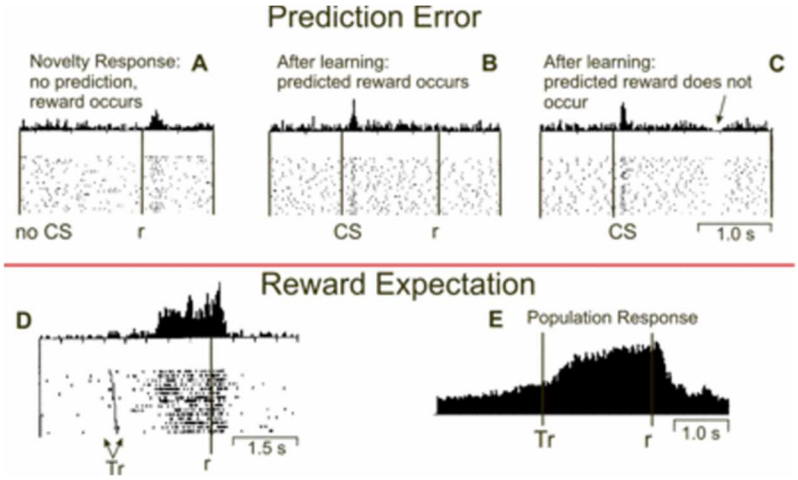


Figure 1.9. Examples of a Prediction Error (*pe*, A-C) and some Reward Expectation (*re*, D, E) neurons [27]

However, only few dopaminergic neurons produce error signals that comply with the demands of reinforcement learning. Most dopaminergic cells seem to be tuned to arousal, novelty, attention or even intention and possibly other driving forces for animal behavior. Furthermore, the TD-rule reflects a well-defined mathematical formalism that demands precise timing and duration of the δ error, which cannot be find in neural structures. Consequently, it might be difficult to calculate predictions of future rewards. For that reason alternative mechanisms have been proposed which either do not rely on explicit predictions (derivatives) but rather on a Hebbian association between reward and CS [28], or which use the DA signal as a switching signal which establishes learning after salient stimuli [29, 30].

Differential Hebbian learning seem to be, to some extent, compatible with novel findings on spike-timing dependent synaptic plasticity (STDP) [32]. In this type of plasticity, synapses potentiate (become stronger) when the presynaptic input is followed by post-synaptic spiking activity, while else they are depressed (become weaker). We will envision now a more biologically plausible model for the Critic, including the learning of values.

1.6.3.1 PVLV model

Given that there are distinct brain areas involved in these different aspects of the dopamine firing, it raises the question as to how the seemingly unified TD learning algorithm could be implemented across such different brain areas. In response to this basic question, the PVLV model of dopamine firing was developed[28]. PVLV stands for *Primary Value Learned Value*, and the key idea is that different brain structures are involved at the time when primary values are being experienced, versus when conditioned stimuli (learned values) are being experienced. This then requires a different mathematical formulation, as compared to TD.

More generally, the unitary nature of the TD framework does not seem compatible with the relatively large and diverse cast of brain areas involved in driving dopamine firing. In contrast to TD, PVLV predicts that it should be possible to doubly-dissociate the CS and US associated DA firing behavior empirically.

More precisely the PVLV model separates the acts for the CS stimulus of producing the dopamine spike (learned

reward) from that of shunting the US-onset (reward) dopamine dip (innate reward prediction). Therefore, the model is made up of two systems:

- a system for learning of rewards , called *LV (Learned Values)*
- a system which learns to predict innate rewards called *PV (Primary values*)

Both systems have an excitatory (PV_e, LV_e) and inhibitory component (PV_i, LV_i) and learn according to the simple Pavlovian conditioning paradigm.

The model explored, derives from a more recent implementation [35], which is shown in Figure 1.10.

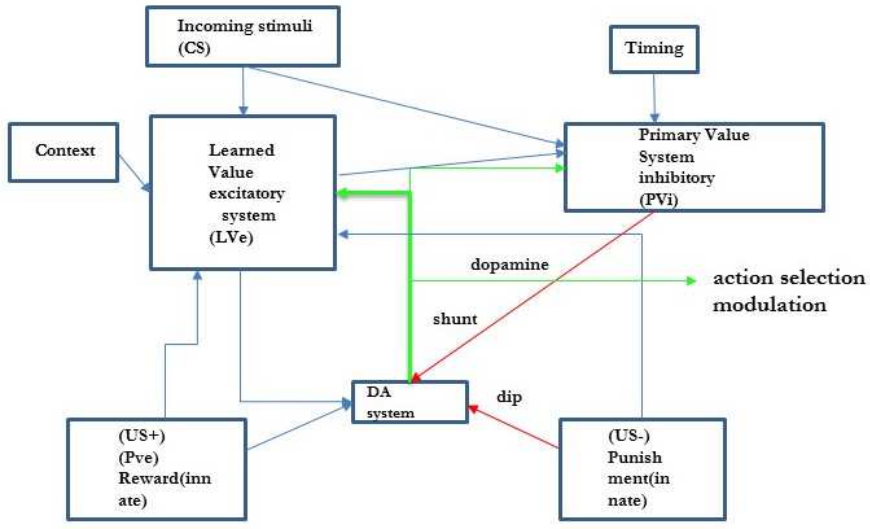


Figure 1.10. The PVLV framework learning occurs in parallel in two systems in a Pavlovian paradigm

The model takes in account primary rewards (US+) and punishments (US-). Both these constitute the Primary Value excitatory (PV_e). The Leaned Value block learns to produce Dopamine firing on the onset of incoming CS stimuli. We also notice that dopamine self-reinforce the system. Primary Value inhibitory system instead learns to predict innate rewards by using a representation of time (timing block) as input together with the CS. PV_i block act also like LV_i for prediction of learned rewards in second order conditioning paradigms. The context block provides information about the environment in which CS is effective as a reinforcer (motivating operations). The main hypothesis on which the model lies are:

- PV_e always produce dopamine burst (or dip for punishment);
- Learning in the LV_e system occurs only on the onset of US stimulus;
- CS onset dopamine burst cannot self-reinforce.

According to this hypothesis, learning in the two systems occurs according to the rules:

For PV_i:

$$\Delta w_i^t = \in (PV_e^t - PV_i^t)x_i^t$$

For LV_e:

If $PV_e^t > 0.8$ *or* $PV_i^t > 0.8$) *or* ($PV_e^t < 0.2$ *or* $PV_i^t < 0.2$)

$$\Delta w_i^t = \in (PV_e^t - LV_e^t)x_i^t$$

Else

$$\Delta w_i^t = 0$$

While the delta rule is:

If $(PV_e^t > 0.8$ *or* $PV_i^t > 0.8)$ *or* $(PV_e^t < 0.2$ *or* $PV_i^t < 0.2)$

$$\delta = \begin{cases} \delta_{pv} = PV_e^t - PV_i^t \\ \delta_{lv} = LV_e^t - LV_i^t \end{cases}$$

Figure 1.11 depicts the output of the DA block in different phases of the learning process.

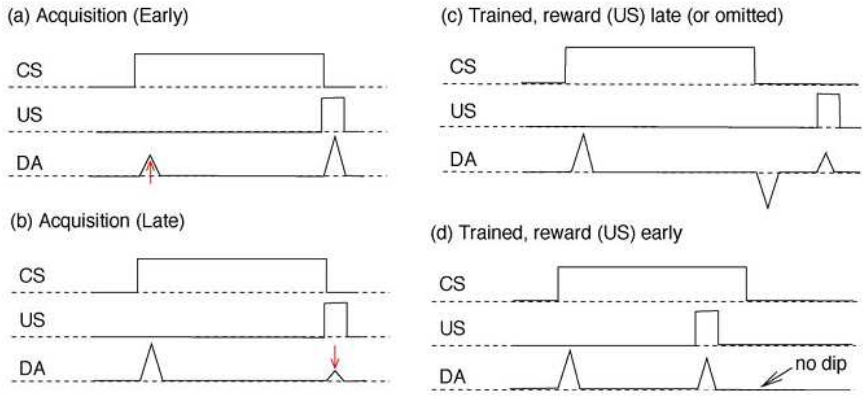


Figure 1.11. The PVLV learning process in various phases of learning: a) early stage of learning: only US can elicit DA firing b) late stage of learning: CS is able to elicit DA firing in place of US; c) in a trained system, in case of wrong prevision (US occurring late or not) a dopamine dip happens when CS activity goes to baseline

d) in a trained system, in case of right prevision (US occurring early) another Dopamine spike occurs on US onset

We can make further considerations and comparisons between TD and PVLV. TD can be thought of as a delta-rule learning mechanism that straddles two adjacent points in time, with the present serving as a training signal for the immediate past. In contrast, the Rescorla– Wagner version of the delta rule operates within a single time during which both an expectation and US outcome are encoded (active) and compared. Thus, at a mathematical level, TD and PVLV (which can be thought of as an elaborated version of Rescorla–Wagner) share the basic delta-rule dynamic, but they differ principally in how they use time for computing the DA delta.

This difference can be captured to some extent within the TD framework itself, with a time-smearing mechanism (‘‘*eligibility traces’’*) in the TD (λ) formulation [31]. The λ parameter, as already mentioned, controls the exponential rate of decay of prior stimulus representations which are available to the learning mechanism, with (λ=0) having no such trace at all, and (λ=1) having an infinite trace. Even with the addition of eligibility traces to the TD (λ) framework, however, other significant issues for TD as a biological model remain. First, TD (λ) does not make a very clear distinction between CSs that persist through to the time of reward (delay conditioning), and those that do not (trace conditioning): the exponential trace enables similar learning to take place in either case.

The standard account of this dissociation is that additional brain systems are necessary for actively maintaining a neuronal representation of the stimulus through to the point of

reward, bridging the gap so that an association can be established. It is unclear why these separate memory systems would be required within the unitary TD (λ) framework, especially because a large λ is generally required to account for even the delay conditioning data.

In contrast, PVLV makes a clear distinction between trace and delay paradigms, because it can only learn about a CS at the time of the US. If the CS is no longer present, some other internal representation of it must be preserved to bridge the associative gap.

For TD (0), i.e., no temporal smearing, each moment in time is completely distinct, every individual time step is indispensable, and temporal chaining is essential for bridging over delays. On the other hand, as one approaches the other extreme of TD (1), chaining becomes much less important, and the system converges in some ways back to the simpler Rescorla–Wagner temporal behavior.

A longstanding empirical issue for TD (0) has been that phasic dopamine firing ought to be seen chaining backward in a 'bucket brigade-like' manner during the CS–US *inter stimulus interval* (ISI) early in training. This is because the way that TD works is to pass the prediction error signal backward one time step at a time with each training trial, implying that phasic DA firing ought to be seen between the CS-onset and the US. This behavior has generally not been observed empirically, but it has been suggested that evidence for chaining may be buried in the noisy ISI period.Instead, firing is generally interpreted as jumping directly to the time of CS-onset by most authors. Specifically addressing this issue, there have been presented new dopamine firing data that largely replicated earlier patterns of firing, and then showed that only the TD (l) model, but not the TD (0), was able to

reasonably simulate the empirical data. In particular, large values (close to 1) were required, such that learning from one time step almost completely generalizes to all previous time steps, thereby almost reducing TD to just the standard Rescorla–Wagner rule.

Finally, the restriction that the LV system cannot reinforce itself, which has clear computational motivations as described above, gives rise to another important difference between TD and PVLV in the context of second and higher order conditioning. Because a CS elicited DA burst would not be expected to train further CS-CS associations within the PVLV system, this mechanism should not support higher order conditioning. In contrast, the unitary nature of TD causes it to automatically and easily support arbitrarily high orders of conditioning. As pointed out in the original paper [28], there is a dearth of evidence for higher order conditioning beyond second order conditioning. While the absence of direct evidence is clearly not evidence for absence, we suggest that even second order conditioning displays characteristics quite distinct from first order, in a pattern suggesting it may not be primarily dependent on phasic DA firing .

In summary, both PVLV and TD (λ) models (with λ closer to 1than to 0) make many of the same predictions that are consistent with known empirical phenomena, because both use a similar treatment of time, and both are fundamentally delta-rule/Rescorla–Wagner based mechanisms. However, PVLV also makes other predictions that TD does not make, which also seem consistent with available data. Some researchers will likely prefer the theoretical elegance of the TD framework, particularly as a normative model in the context of traditional artificial intelligence

In the following chapter, we will explore the neural correlates of reinforcement learning.

Chapter 2

Neural substrates of reinforcement learning

In the previous chapter, mechanisms of learning were explored, showing how from the innate reflex to the most sophisticated form of voluntary behavior learning occurs. We provided the basics of reward learning and reward based learning rules. We also explored how a system of innate and learned rewards exists. Now we will use this concept to build and explain a plausible model of reward based motor learning, exploring in depth the neural correlates structures of this model. We will discuss how the system can be seen as made up of layers structured in a hierarchical fashion, as in the previous chapter we found out how high level behaviors originate from five main questions (*H4W problem*).

2.1 A hierarchical model

To build this hierarchical model, we consider three main factors: exteroception (*world*), interoception (*self*), and action.

At first, the brain must assets the motivational states derived from homeostatic self-essential variables, the *needs*. These motivational states in turn need to be prioritized so that *goals* can be set: this is the *Why* problem, requiring modulation of underlying behavior system.

Next, a layer of control is called for to classify, categorize and evaluate states of the world, to identify the spatial layout of the task, including the agent itself, and the dynamic of the task and its affordances :*' What , Where, When'*.

Lastly, these labeled multi-modal states are grouped in sequences around prioritized goals. Using the accumulated spatiotemporal knowledge of the task and the self in which the goal pursuit is framed, a procedural *motor strategy* can be composed and its elements selected from the set of available options to achieve a goal state: *How*.

Starting from the lowest level the system can be seen as made up of hierarchical levels:

- A first level which designates the body itself and defines three fundamental processes: exosensing of states of the environment (*sensors*), end sensing of states of the body or essential homeostatic variables of survival defining needs (*special sensors*) and actuation through control of the skeletal-muscle system (*motor outputs*). Behavior is defined as a change in the confirmation and/or position of the somatic level, and consists of limited and stereotyped hardwired stimulus-response associations, *reflexes*.
- A second level comprises dedicated *behavior* systems that combine predefined sensorimotor mappings (*reflexes*) in order to reduce needs of the body. A need

arises from the discrepancy between a read-out of a homeostatic parameter (e.g. blood sugar level) and an optimal set point. This system so learns new reflexes trough classical conditioning paradigm. Further, this level modulates the engagement of higher control layer and their epistemic (cognitive) functions (for example, making a movement smoother).

- A third level extends the predefined need-reducing sensorimotor loops of second level with *value-dependent acquired sensor and action states*. Given the definition of goal-directedness and the limitations of primary needs systems in this respect owing to their reliance on fixed action patterns, there is a need for 'higher' systems to be informed about drives expressed by 'lower' levels. The acquired sensor and motor states are in turn associated through the values states triggered by the second level, following the paradigm of classical conditioning where initially neutral or conditioned stimuli (CS) obtain the ability to trigger actions, or conditioned responses (CR), by virtue of their contingent presentation with intrinsically motivational stimuli or unconditioned stimuli (US), (*innate rewards*). Voluntary movements are selected according to a policy, which get strengthened/weakened by reward/punishers. A *context* **(***motivating operations***),** as seen in the previous chapter, is fundamental for giving a reinforcer/punisher a salience, so context deeply shapes the form of a voluntary movement.

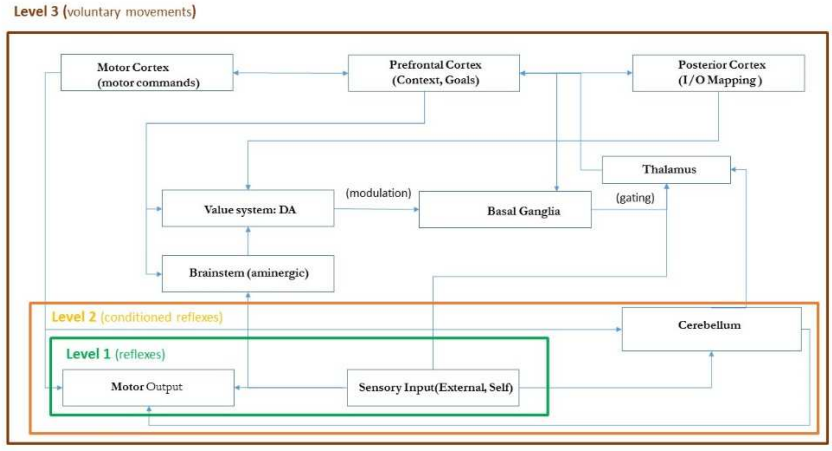


Figure 2.1. A hierarchical structure emerging for learning of voluntary movements, appears to be composed of three levels: spinal reflexes (green), conditioned reflex (orange), voluntary movements (brown)

Regarding neural correlates of the system, we can start making some statements:

- The first level involves Motor Output and Sensory input,
- The second level includes Motor Output, Sensory input and Cerebellum
- The third level includes the sub-systems of action selection, trough the Basal Ganglia (*How*) and a Value System (*Why*).
- The Cortex fulfils the role of collecting high level representations of perception of the world coming from the sensory (thus answering the *Where*, *What* and *When*) and motor plans/action to pursue, going to motor actions, as well providing a context and goals of

the task. The presence of Prefrontal Cortex and *aminergic systems* (DRN) provides motivation to the system, creating the context, realizing so the *four term contingency*.

We will go more in detail with point 3 as the first 2 points are beyond purpose of this thesis.

2.2. Learning First level: spinal reflexes

At the lowest level, the movements as mentioned before essentially consist in *myotatic reflexes* [35, 36]: they make possible to keep a group of muscles in tension in order to maintain body at a given resting position. The neural correlate of this layer is the spinal cord, whose H-shaped section involves a direct association of sensory inputs with motor outputs. Modulation of such reflexes involves *gamma circuits* [37].

Figure 2.2 A reflex is a stimulus response association, occurring trough gamma circuits in the spinal cord

Gamma motor-neurons in the ventral part of the s*pinal cord* regulate the level of activity of *the intra-fusal fibers* thus setting the resting position; the myotatic reflex will regulate the *alpha-motor neurons activit*y, activating *extra-fusal fibers* in such a way that the force applied by the whole muscle keeps the position set by gamma motor-neurons. Information about mechanical action performed is coded by the *muscle spindles* carrying information about position/velocity (*proprioception*).

The information coded by the muscle spindles are also sent to the *cerebellum* as *mossy fibers* and to the *cerebral cortex* via the *thalamus*. The activation of a myotatic reflex, namely a strong/abrupt change in alpha motor neurons, is also signalled to the *inferior olivary nucleus*.

2.2.1 Learning of movements in human brain: from reflexes to voluntary actions

Primary reflexes are still present in a "spinal animal", that is when there is a lesion that cuts the contiguity between the brain and the spinal cord, meaning that the cerebral cortex, the basal ganglia and the cerebellum can no more take place in movement control [38]. This demonstrates that primary reflexes are due to spinal circuits and to the equivalent *medulla circuits*. The difference between spinal cord and medulla is just in the parts of the body that they control: trunk, limbs and viscera for the spinal cord; head, neck and viscera for the medulla. In a spinal animal, there are no voluntary movements and no conditioning; this demonstrates those voluntary movements (operant behavior) and learning (operant conditioning) involve other nervous centers. Actually, some forms of rudimental learning are still present in a spinal animal. The most studied is called "*conditioning of the H-reflex*"; this mechanism can be responsible for some kind low-level modulation of the spinal reflexes [39].

2.3 Second level: conditioned reflexes

The neural substrate for this level seems to be Cerebellum. The role of the Cerebellum and its associated circuitry in the acquisition and retention of anticipatory responses (sensory

predictions) with Pavlovian delay conditioning has been well established [40, 41]. Although most of the classical conditioning studies are primarily based on eye-blink conditioning [42], recent experimental studies have established the essential role of the cerebellum in learning and memory of goal-directed behavioral responses [43]
Learning in classical conditioning regards sensory prediction. As the Rescorla–Wagner model formalized in the previous chapter, animals learn in classical conditioning only when events violate their expectations [44].

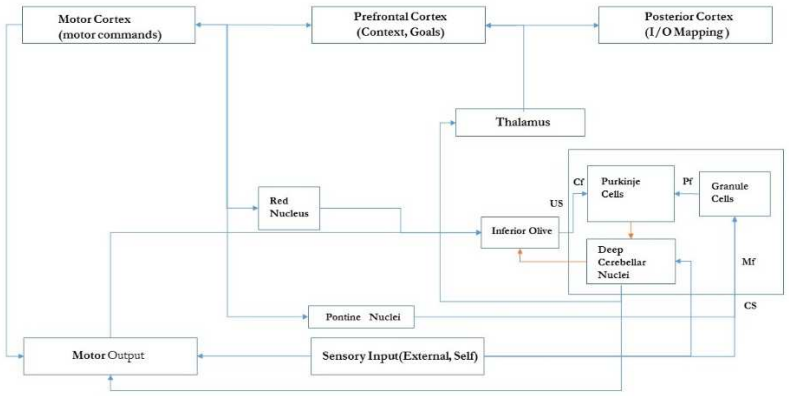


Figure 2.3. The cerebellum, through the association of its two inputs, is able to create conditioned reflexes from lower level, yet, trough thalamic projections, contributes to modulate voluntary movements from higher level

The CS and the US in Cerebellum comes respectively from its two inputs, the *mossy fibers* and the *climbing fibers*, and the motor response output CR comes from *deep cerebellar nuclei*.

The US signal relayed by the Inferior Olive (IO) reaches the cerebellar cortex through the climbing fibers where it induces plasticity at the synapses of the parallel fibers that transmit the CS information. After repeated coincidence of these two signals, the *Purkinje cells* – the sole output of the cerebellar cortex – acquire a response to the CS, namely, a drop in their firing activity that drives the behavioral CR.

At first, mossy fibers are unable to activate deep nuclei neurons, due to the inhibitory action opposed by Purkinje cells, which are also activated by the same mossy fibers through the *granule cells*.
Whenever an association occurs, i.e. a concomitance of a (non effective) stimulus and the activation of a reflex, it happens that the synapse between the granule cells activated by that stimulus and the relative Purkinje Cell is "cut off", a phenomena called long term depression (LTD) [43]; from now on, that stimulus will activate some deep nuclei cells, because the same stimulus does not activate the inhibitory branch (Purkinje cells). At the same time, the synapse between the parallel fiber and the Purkinje cell is potentiated trough long-term potentiation (LTP) [44].
Given that Granule cells code combination of stimuli and Deep Cerebellar nuclei code combinations of motor action, Cerebellum is able to match sensory inputs with motor actions creating new conditioned reflexes [45]. This can be a way to react to environment (like avoiding a danger) and in this case the CS input comes from low level sensory input (spindles), while the response goes to motor neurons (alpha motor neurons) [46]. Through Hebbian mechanism, synapses strengthen among the cerebellar nuclei neurons that activate and the motor neurons that are active at the same time, so in

the end a new reflex arc is created between new sensitive stimuli and the motor neurons activated simultaneously with these sensitive stimuli
According to [47], cerebellum only learns when the IO activity is perturbed from baseline. In this context, the inhibitory connections from the cerebellar deep nuclear cells to the Inferior Olive, the *Nucleo Olivary Inhibition* (NOI) [48], are key to interpret cerebellar learning as the acquisition of sensory predictions. The NOI subtracts the cerebellar output relayed by the deep nuclei from the US signal reaching the IO, such that if both the signals match, they cancel each other leaving IO activity at baseline [49]. IO olive so acts by giving a sort of teaching signal.

In order to make IO signal comparable with the inhibiting signal from Deep Nuclei, which codes an action, we assume that it also codes an action, not a sensory state, precisely an abrupt change in the position maintained by the myotatic reflex [49]. In case of voluntary movements, the teaching signal of the IO is a high-level cognitive cue coming from the cortex trough *Red Nucleus* [50].

Cerebellum also contributes to make a movement smooth [51], like, for example, the act of writing a letter without the need to correct through voluntary movements all the strokes that made for it. In this case, the input CS comes from the cortex trough pontine nuclei while the output goes upstairs through thalamus to the Cortex , summing up to the output of Basal Ganglia, so modulating the action of the voluntary movement. At the beginning of learning, voluntary circuits recognize “errors” and apply corrections [52]; by repeating these voluntary corrections many times under the same circumstances (i.e. the same incoming stimuli), the

cerebellum recognizes the simultaneous occurrence of these stimuli and movements and is solicited to link them in a new, conditioned reflexes. The cerebellum is solicited to do this because the activation of movements means also the activation of the inferior olive.

2.3.1 Neurophysiological considerations

The ability to acquire conditioned reflexes is preserved if there is a lesion of the brain cortex and the basal ganglia that spares the cerebellum; under such conditions, the animal stays still and voluntary movements are absent. This demonstrates that the cerebellum is responsible for the acquisition of conditioned reflexes but not to the initiation of voluntary movements. One of the most used model to study the functioning of the cerebellar circuits is the eye-blink conditioning. The acquisition of a new, conditioned reflex happens when there is a stimulation of the inferior olive [39]: this lead to the formation of a sort of link between those stimulus and movement that were contingent with the stimulation of the inferior olive. Signals to the inferior olive may originate from the spinal cord or from the *pontine nuclei*. Pontine nuclei may be activated, in turn, by the cerebral cortex.

In this thesis, as mentioned in the previous section, it is postulated that the signal from the spinal cord to the inferior olive originates from the activation of a reflex; this hypothesis is supported by the experimental finding that the olivary neurons integrate information pertaining to individual spinal

withdrawal reflex modules, and to muscles of functionally related modules [49]. Thus, it has been postulated that whenever an effective stimulus *A* evokes a reflex, the cerebellum is prepared to link to the same motor response any other stimulus *B* that is contingent to that stimulus *A*.

By this way, a new stimulus *B* becomes able to evoke that particular reflex that was activated (the first times) by the effective stimulus *A*.

2.4 Third Level: voluntary movements

The third level makes use of the lower levels in order to achieve adaptability in learning voluntary movements. As exposed in the previous chapter, the probability of a voluntary movement occurs according a 3-terms contingency (4 considering motivating operations). The *Actor-Critic* is one of the most accredited architecture of operant conditioning and is made up of two subsystems working in parallel. In the previous chapter, we introduced the Actor –Critic architecture and its function, now we will focus on the neural correlates.

From a computational perspective, the key idea is the distinction between an *actor* and a *critic* where it is assumed that rewards result at least in part from correct performance by the actor.

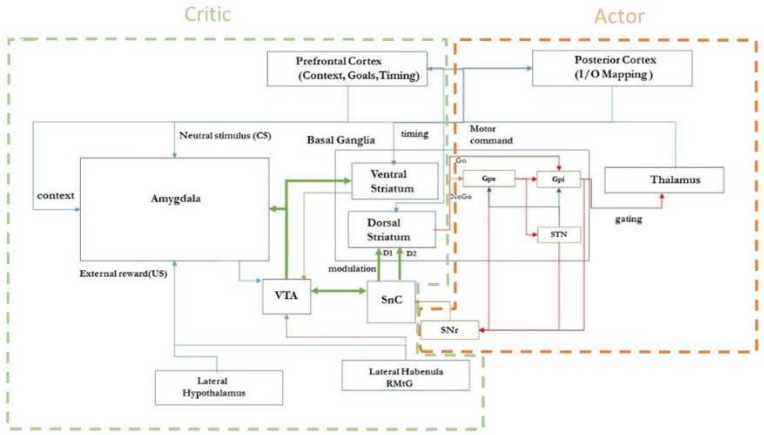


Figure 2.4. The higher level for generation of voluntary movements relies highly on an Actor- Critic structure, where the Critic system (green dotted line) is able to modulate the Actor (orange dotted line) actions through dopamine signalling, which could be triggered by innate and learned rewards/punishers.

The dopamine signal is the output of the Critic, which then serves as a training signal for the Actor (and the Critic too, as we saw in the previous chapter). The reward prediction error signal produced by the dopamine system is a good training signal because it drives stronger learning early in a skill acquisition process, when rewards are more unpredictable, and reduces learning as the skill is perfected, and rewards are thus more predictable. If the system instead learned directly on the basis of external rewards, it would continue to learn about skills that have long been mastered, and this would

likely lead to a number of bad consequences (synaptic weights growing ever stronger, interference with other newer learning, etc.). In the following sections, we will go in detail with this model, which comprises:

- the Critic, made up of the DA system, which modulates the action selecting mechanism in the Actor system:
- the Actor, made up of Cortex and Basal Ganglia

2.4.1 The Critic

We will start reviewing the neural structures, which made up for the Critic system. As already discussed, the role of this system is to produce a teaching signal for the Actor. This signal consists in a dopamine firing which codes some sort of reward prediction error between a reward (punishment) and an expectation of a reward. This dopamine firing occurs in the *Ventral Tegmental Area* (VTA) and *Substantia Nigra part Compacta* (SnC), nuclei located in the midbrain, which appear to have the biggest concentration of dopaminergic neurons.

We also found out that rewards and punishment can originate from innate values (external reward) or an initially neutral stimulus, which acquires rewarding abilities after a Pavlovian conditioning. External rewards are signalled mostly by *Later Hypothalamus*, while punishment by *Lateral Habenula*. It will be shown that the neural substrate for this conditioning occurs in the amygdala and Striatum (part of the Basal Ganglia) for the prevision of rewards. Figure 2.5 depicts the architecture of the Critic system

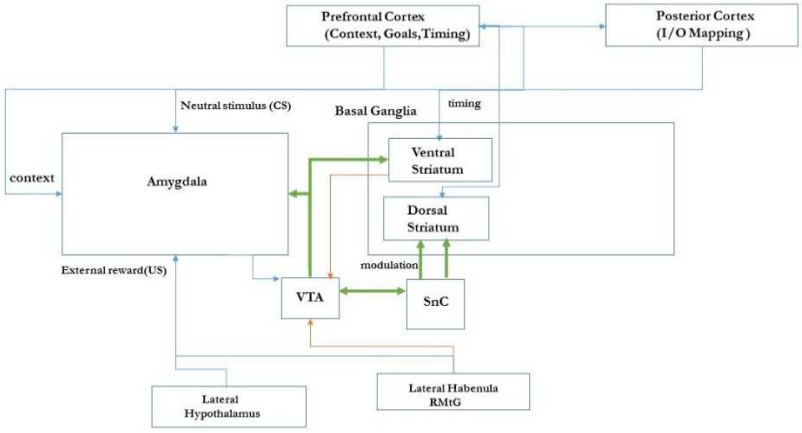


Figure 2.5. The Critic system consists in many structures contributing to produce dopamine, encoding a reward predicition error. Rewards and punishment could be innate or learned.

2.4.1.1 Lateral Hypothalamus

Perhaps the most well-known brain region for controlling motivational drives is the *hypothalamus*, a phylogenetically ancient diencephalic structure well connected to sensor and actuator systems in lower CNS centers such as the brain stem, spinal cord and autonomic ganglia. Hypothalamic cell groups are thought to monitor nutrient's levels and guard the body's energy balance, while others regulate sexual, maternal and aggressive behavior as well as sleep. This list of hypothalamic sensor –actuator functions is by no means exhaustive and can

be supplemented with numerous brain stem-medulla nuclei that are often positioned even closer to internal sensors and effectors (e.g. the monitoring and regulation of food intake, respiratory and cardiovascular reflexes by the *nucleus tractus solitarius*, *vagal nuclei* and connected cell groups). In addition, these functions address not only homeostatic regulation, but also *allostasis*, referring especially to responses to challenges that require system-wide, dynamic adaptation and predictive regulation in anticipation of upcoming homeostatic disturbances [53, 54].

The Lateral Hypothalamic Area (LHA) is responsible for reactive representation of US value for rewarding stimuli such as food, water, etc. and this provides the main excitatory signal driving phasic dopamine bursting after primary reward onset. This is a widely accepted hypothesis, as cells of LHA receive direct projections from primitive sensory areas associated with primary reward [55,56] and respond to the occurrence of a reward with *sustained* firing [57,58,59,60].

The LHA sends excitatory glutammaergic projections directly to the midbrain dopaminergic nuclei (VTA and SNC, [61]) and even more densely to *the pedunculopontine tegmental nucleus* [62], which in turn sends both glutammaergic and cholinergic projection to midbrain DA nuclei [63]. Further, it has been shown that the strength of DA synapses is relatively fixed, hardwired during ontogeny or VTA has little plasticity [64, 65, 66]. This is in accordance with the assumption made in the former chapter that the Primary Values system (*PVe*) does not learn so that an US will always elicit dopamine firing.

2.4.1.2 Lateral Habenula

The habenular nuclei are involved in pain processing, reproductive behavior, nutrition, sleep-wake cycles, stress responses, and learning [66]. Recent demonstrations using *fMRI* [67] and *single unit electrophysiology* have closely linked the function of the lateral habenula with reward processing, in particular with regard to encoding negative feedback or negative rewards. The authors of [67] suggested that this reward and reward-negative information in the brain might "be elaborated through the interplay among the lateral habenula, the basal ganglia, and *monoaminergic (dopaminergic and serotonergic) systems*" and that the lateral habenula may play a pivotal role in this "integrative function".

The *rostromedial tegmental nucleus* (RMTg), also known as the tail of the ventral tegmental area (VTA), is a GABAergic nucleus, which functions as a "master brake" for the midbrain dopamine system [68, 69]. It is poorly differentiated from the rest of the ventral tegmental area (VTA) and possesses robust functional and structural links to the dopamine pathways. Recent findings that the lateral habenular nucleus can produce pauses in DA cell firing [71, 72, 73] suggest that it may be important for generating the dip associated with omitted rewards.

2.4.1.3 Amygdala

The amygdala proves to be the main brain structure responsible for making associations of different predictive

cues with primary affective outcomes [74]. It is generally considered one of the most important regions of the limbic system, or brain emotional centers. At a rough functional level, the amygdala can be divided into a cortex-like *basolateral set of nuclei* (BLA; basal, lateral, accessory basal nuclei), and a more striatum-like *central set of nuclei* (CNA; medial segment, lateral segment) [75, 76]. Both BLA and CNA receive broad projections from all over the cortex, with the CNA receiving such projections both directly, and via a kind of funneling pattern from the BLA. *Multimodal cells of medial segment* of the CNA (mCNA) sends some excitatory projections directly to the midbrain DA nuclei [77, 78], and to the LHA [79]. In contrast, the BLA does not project independently to the DA midbrain areas, only indirectly doing so via its projections to the mCNA. Finally, electrophysiological studies have provided strong corroboratory evidence that stimulation of CNA neurons can cause dopamine cell firing in the VTA and SNc and/or DA release in target areas [80, 77].

While the amygdala has long been associated with fear conditioning (see, [81] for a review), it is now well established that both the BLA and CNA also code for positively valence stimuli [82, 83, 84, 74]. mCNA are initially responsive only to US, but subsequently these cells respond also to a CS paired to the US [74]. As a result, CS-onset also acquires the ability to drive DA bursting via excitatory projections from the mCNA to the midbrain DA system (figure 2.6). This crucial link serves to ensure that population of DA cells driven by LVe system (CC-onset) will be approximately the same population driven a priori by PVe system (US-onset).

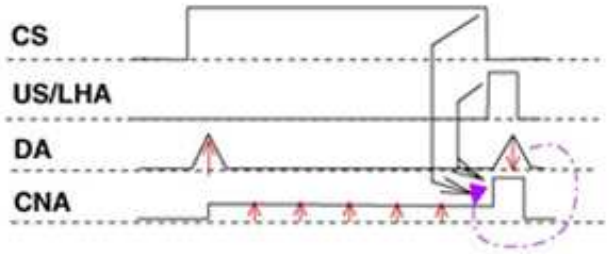


Figure 2.6. The neural substrate for the acquisition of learned rewards appears to be central nucleus of Amygdala (CNA), which exhibits a sustained firing pattern during the duration

Furthermore, consistent with the idea that L-LTP (late, or permanent, *long term potentiation*) is occurring for these events, immediate early gene expression has been observed in CNA cells, particularly those that project to SNc, in response to a visual stimulus predictive of reward [85].

The final evidence for the CNA playing the critical role in driving phasic DA bursting at CS-onset comes from studies showing that CNA lesions interfere with a set of Pavlovian conditioned responses that are likely to depend on CS-driven phasic DA.

According to a more elaborated representation of the amygdala circuitry [86] Basal Amygdala comprises different portions able to create learned positive values, and learned negative values (*Basal Amygdala for Acquisition of positive/ negative values*). It also comprise portions that learn on the omission of expected rewards/punishments (*Basal Amygdala*

for extinction of positive/negative values). In the case of extinction of a reward, an important role is that of context, which stands for motivating operations and is provided from Orbitofrontal Cortes, as it will be showed in the following sections.

2.4.1.4 Ventral Striatum

Separately and in parallel with the learning of new values, a system comes to expect the US by learning about the system's internal state, including temporal representations, immediately prior to the US onset. This acquired representation then acts via GABAergic projections (and shunting inhibition) to 'cancel' the DA spike at the time of reward. In the PVLV paradigm explored in the previous chapter, we called this system PV_i.

The neural substrate for such representations appears to be *Ventral Striatum (Nucleus Accumbens, NAc)*. It is part of the basal ganglia, which includes the *dorsal* (specialized for actions) *and ventral striatum* (specialized for reward information).

Ventral Striatum encodes a 'value' signal incorporating different costs like effort, possibility of pain and risk into an estimate of how good something is. Additionally, by projecting to the dorsal striatum, it can use these value signals to help take actions that lead to possible rewards. The ventral striatum also has two distinct sub-populations of *medium spiny neurons*, which have been described as *patch-like and matrix-like* because of the histological staining characteristics they share with their dorsal counterparts [87].

However, the Ventral Striatum (especially NAc) does not exhibit the histological compartmentalization of these cell types seen in the dorsal striatum, which has made it difficult to establish connectivity differences between these cell types similar to those established for the dorsal striatum. Nonetheless, two different subpopulations of *MSNs(medium spiny neurons)* have been described in the NAc based on connectivity, one projecting onto DA cells of both the VTA and SNc and the other synapsing onto GABAergic neurons of the SNr [89,90,87].

Patch-like GABAergic neurons in the ventral striatum (striosomes) are the main substrate for the learned representation of a US expectation (PVi in the PVLV system or value function in the TD algorithm)

In response to a predictive CS, striatal cells are known to acquire *ramping activation* that peaks at the time US is expected [91]. By virtue of inhibitory projections onto DA cells, these acquired representations can then shunt the excitatory signal when the US occurs, eliminating the DA burst previously seen when the US was unexpected.

Inhibitory projections therefrom to midbrain DA cells shunt excitatory inputs thereby eliminating the phasic burst for the US. The VS also projects via to the lateral habenular nucleus of the *epithalamus*, helping that substrate to compute when an expected reward has been omitted. During conditioning, midbrain dopamine neurons eventually stop firing a phasic burst at the time of a reward, and if an expected reward is then omitted (or delayed), there is a phasic pause or dip in tonic firing at the expected time(figure 2.7). These effects are thought to be global across the majority of DA neurons in the VTA and SnC [92].

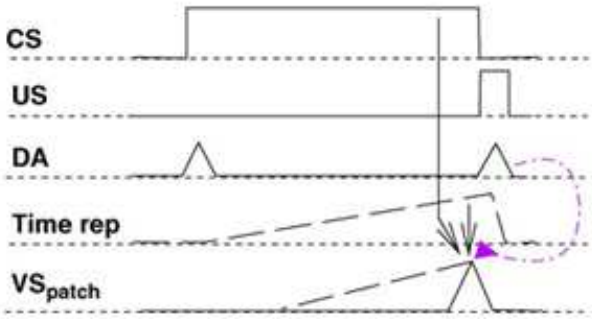


Figure 2.7. Patch-like neurons in the Ventral Striatum exhibit a ramping activity such that they are able to cancel the excitatory signal to DA cells. In response to a predictive CS, a subpopulation of ventral striatal, the ramping cells in the Ventral striatum seem to be integrating timing information in order to peak at the time of anticipated US.

As regards the ability to predict CS-onset DA firing, in second order conditioning (a learned reward is associated with another neutral stimulus) the mechanism occurs with a slower rate (called LV_i in the PVLV system). Schultz et al. [92] found cells in the VS that fit this exact pattern: subpopulations of VS neurons exhibit peaking at the CS-onset [92, 93], triggered by the occurrence of a still prior stimulus. In the study, the prior CS 2(second conditioned stimulus) was explicit, but it is easy to imagine, that animals could develop LVi representations from implicit/contextual signals as well such predictions would be less exact and, therefore, the mitigation of dopamine firing only partial, which is what is seen empirically with overtraining and no explicit CS2 [94].

The source of *timing representation* in striatum remains an empirical question. One obvious candidate is the cerebellum, widely thought to be important for representing time [95]. This would require cerebellar input to influence VS neurons, which does not happen through direct projections, but there are indirect pathways via *cerebellothalamic* and *thalamostriatal projections* [96], and via the frontal cortex [97]. Another candidate could be Orbitofrontal Cortex, part of the Prefrontal Cortex [98]. Another proposal is that striosomes can actually exhibit timing dynamics themselves [99].

2.4.2 The Actor

The roles of the Cortex and of Basal Ganglia seem to fit well with the architectural hypothesis of an Actor, a system which is able to increase the frequency of an action (*policy*), given an ensemble of stimuli inputs. In this scenario, the Cortex is responsible for storing all possible actions and *sensory combination* (posterior), as well as to initiate *motor actions* (motor, prefrontal).

The basal ganglia acting like a "switch" enable or prevent a possible behaviour from being executed, according to its probability. Motor activations coming from the cortex translate, at "low level", in modulations of reflex arcs, i.e. modulations of the activity of gamma motor neurons.

Whenever a dopamine firing occurs (reward or punishment), at the cortical level there will be an increase (or decrease) in the probability associated with the stimulus-action pairing. Figure 2.8 shows the model for Actor-Critic.

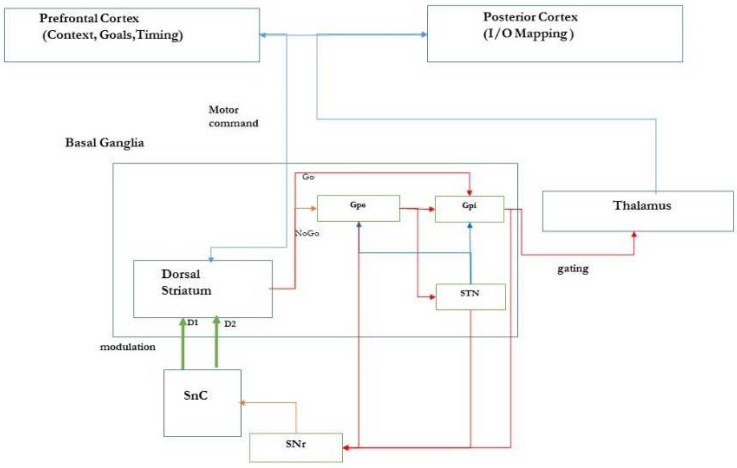


Figure 2.8 The Actor can select an action by selecting Go or NoGo pathways for that behavior according to DA firing signaled by SnC in Dorsal Striatum D1 and D2 receptors

2.4.2.1 Cortex

The cerebral cortex can be viewed as a collection of information that can be extracted from the incoming stimuli (sensory cortex, posterior), and repertoire of possible movements that can be initiated (motor cortex, anterior).

Within the cerebral cortex there is a large amount of links between the sensory and the motor cortex, so that given a stimulus there are several different possible movements that can be initiated, but with different probabilities.

Structures such as *posterior parietal cortex* and *primary sensory cortex* (A1, V1) are deeply affected by associative stimulus–reward and action–reward learning [100-105]. Despite this ubiquity, there are good arguments to highlight

the role of prefrontal systems in forming *state representations* that can be used by *action selection systems* executing *goal-directed behavior* [106]. Put succinctly, when the needs of an agent *('Why')* have been set at the level of the hypothalamus and brain stem, representations of the state of the world (including the agent's own state) are required to determine *Where* and *When* this need may be satisfied, and through which particular object *('What')* within a feasible spatio-temporal goals (e.g. an apple to satisfy the need for particular nutrients) and goal sites themselves.

Posterior Cortex codes representations of the *state of the world*, required to fulfil a goal. Therefore, a collection of combinations of sensory inputs could be associated to a repertoire of possible actions. The Prefrontal cortex is more concerned with *task- and action- space representations*. Importantly orbitofrontal and medial prefrontal-anterior cingulate neurons are sensitive to motivational value of cues [107,108,109] and actions associated with goal pursuit [110]. It is made up of *stripes*. Thus whereas Posterior Cortex represents high level of perception *(What, When and Where)*, the prefrontal cortex appears better equipped to represent a *task space*, i.e. the set of rules, constraints, goals and goal-predictive values of cues and actions available as options to pursue goals (How) [111].

2.4.2.2 Basal Ganglia

The anatomical architecture, internal wiring and information resources in afferent structures place the basal ganglia in an eminently suitable position to, first, code *state–outcome relationships*, 'state' can be stimulus, place or action, and,

then, to use this associatively learned information to force an expected outcome-dependent decision among response options represented in task space.

The *dorsomedial striatum* has been implied in *action-outcome learning* [112] and as it has been implied in *habit formation* and *sensorimotor learning* with minor or no dependence on motivational outcome. Its role can in fact be very well accommodated if the 'outcome' is viewed more broadly: outcome can also be constituted by actions, so that cue-action (or stimulus–response) learning is subsumed under an overall basal ganglia architecture for 'input –outcome' learning. An essential organizational feature of the basal ganglia is the grouping of topographical projections in parallel 'loops', starting in a particular cortical area and, from there, projecting to specific striatal sectors (dorsal striatum), external segment of the *globus pallidus* and output structures, such as the *substantia nigra reticulata* [113]. By themselves, these loops do not illuminate a specific mechanism for selecting among available response options.

However, striatal principal cells are connected via GABAergic recurrent collaterals, providing a potential mechanism for competitive selection [114,115]. Furthermore, the basal ganglia possess a funnel-like structure in the sense that the downstream flow of processing in cortico-basal ganglia loops is compressed into lesser and lesser neurons.

This structure may provide further competition mechanisms operating at, or in interaction with, the output levels such as substantia nigra reticulata and the internal segment of the globus pallidus [116]. By itself, the presence of GABAergic, inhibitory interactions would suggest an inflexible, learning-insensitive competition mechanism in the striatum. By contrast, recording and pharmacological studies

indicate an active role of the basal ganglia in learning goal-directed behaviors.

The sign of the reward prediction error is appropriate for the effects of dopamine on the Go and NoGo pathways in the striatum. Positive reward prediction errors, when unexpected rewards are received, indicate that the selected action was better than expected, and thus Go firing for that action should be increased in the future. The increased activation produced by dopamine on these Go neurons will have this effect, assuming learning is driven by these activation levels. Conversely, negative reward prediction errors will facilitate NoGo firing, causing the system to avoid that action in the future.

2.4.3 Neurophysiological considerations

In case the cerebellum gets lesioned, but the spinal cord, the cerebral cortex and the basal ganglia are spared, the animal does not stay still, voluntary movements are present, primary reflexes are present, learning by reward is present but the acquisition of new reflexes is not present. Thus, there is a lack in the ability to learn fast responses and accurate movements. All movements have a typical tremor that is the consequence of a long delay between the stimulus (e.g. recognition of a trajectory) and the response (e.g. correction of the trajectory)

This demonstrates that cerebral cortex and basal ganglia are involved in producing voluntary movements and learning by reward. The tremor in such voluntary movements is due to the lack of acquisition of fast responses to stimuli (new conditioned reflexes), which normally take place through the

cerebellum. The cortical responses to stimuli require longer time due to an intense elaboration of the incoming stimuli.

2.5 Neural correlates for motivating operation

In the search for the neural correlates for motivating operations, Balleine et al. [117] argue that there are two different kinds of connections between rewarding/punishing stimuli and the appetitive and aversive affective structures that control performance of CRs: one direct and one indirect via a connection between the CS and sensory properties of the US.

According to the review [118], innate unconditioned motivational operations, like food and water deprivation, have been argued to threshold or gate connections between the sensory US representation and the appetitive system and, therefore, to modulate the indirect link from CSs to the appetitive system. The gate is US specific. It also appears that in other structures involved in ascribing affective significance to CSs, such as the amygdala, neural responsiveness is not gated by motivational state, suggesting that these areas may be involved in the direct pathway associating CSs to the appetitive system. As shown in the previous section, it seems that a context is responsible for the effectiveness of conditioned reward, through a specific connection with Basolateral Amygdala. Correlates for motivating operations seems to be *Dorsal Raphè Nucleus* for the gating mechanism and the *Orbitofrontal Cortex* (Part of Prefrontal Cortex) for context (conditioned motivating operations). The function of cortical/subcortical structures thus regards action selection based on an ensemble of sensory data; this function may have

different degrees of activation, varying from shut down (‘sleeping’), to hyperactive (‘awake’, ’nervous’) according to motivation. Figure 2.9 shows the model including motivating operations correlates.

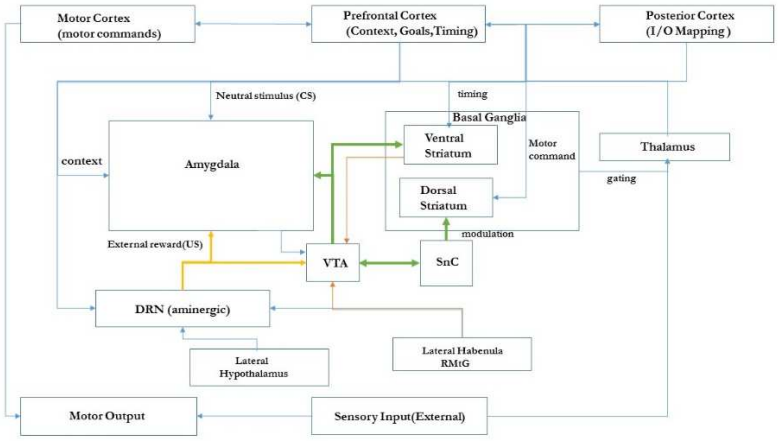


Figure 2.9. Aminerging system (DRN) and Prefrontal cortex provide motivating operations for innate and learned rewards.

2.5.1 Dorsal Raphe Nucleus

The Dorsal Raphe Nucleus (DRN) is part of the Raphe nucleus, which is located in the brainstem. It is the largest serotonergic nucleus and provides a substantial proportion of the *serotonin (5-HT)* innervation to the forebrain. 5-HT has been implicated in a variety of brain functions, such as the sleep-wake cycle, appetite, locomotion, emotion, hormonal

regulation, and as a trophic factor. In addition to the "basic" brain functions described above, the role of 5-HT in cognitive functions, including attention, control of impulsivity, coping with stress, social behavior, value-based decision making, and learning and memory, has also captured a great deal of attention [119]. The breakdown of the 5-HT system is often associated with neuropsychiatric diseases including depression, schizophrenia, drug abuse, autism, and Parkinson's disease.

A growing body of research has revealed that the activity of DRN neurons is related to reward processing. Recordings from primates and rodents have shown that some DRN neurons are sensitive to the expectations, sizes, and deliveries of rewards.

The DRN receives projections from many brain areas that have been associated with reward and punishment. Almost Frontal cortical areas project to the DRN, part of these projections are via GABA interneurons. Subcortical areas projecting to the DRN include the amygdala, hypothalamus, and, most prominently, the lateral habenula nucleus. The dopamine neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) also project to the DRN

Efferent projections from the Raphé nuclei are widespread. Many projection sites include areas that are associated with reward processing, such as the neocortex, nuclei in the basal ganglia, amygdala, hippocampus, and hypothalamus [120]. The positive reward effects of 5-HT have been described mainly in relation to brain self-stimulation experiments where animals perform operant responses such as pressing a bar to receive electrical stimulation of the brain.

The tonic activity of DRN neurons may be ideal to signal a continuous level of motivation and hedonic experience throughout the performance of a task. Such a signal may provide a "reward context" signal to the targets of DRN projections, where the signal may be used differently, depending on the type of 5-HT receptor present.

First, the sustained reward signals in the DRN could be used to track the value of the current behavioral state. Such estimated values have an important role in theories of reinforcement learning, which suggest that the prediction error signal of dopamine neurons is calculated as the difference between the actual and expected reward values. Thus, DRN activity could contribute to the computation of prediction errors by providing the *current state of the expected reward value*.

Second, DRN activity may report the *long-term averaged reward*, rather than *immediate, phasic reward* information [121]. In real life, one needs to integrate flows of information, including both appetitive and aversive events and situations, to achieve better decision making to adapt to external changes. The tonic activation patterns of DRN neurons may be useful in integrating appetitive and aversive information coming from different sources over a substantial temporal span. In the model we built from literature, we simplified the connection by considering the hypothesis of DRN acting as a motivational gate between primary values structures (Later Hypothalamus, Habenula) and dopamine production centres (VTA, Amygdala).

2.5.2 Orbitofrontal Cortex

The Orbitofrontal cortex (OFC) is a prefrontal cortex region in the frontal lobes in the brain which is involved in the cognitive processing of *decision-making* [123]. Previous recordings have revealed that OFC neurons encode predictions of reward outcomes. Electrophysiological recordings from monkeys and rats have shown that OFC neurons encode the *motivational significance of the expected reward* [124, 125]. Moreover, subpopulations of OFC neurons are exquisitely tuned to specific reward features such as temporal delays, spatial directions and reward identities [122]

The DRN densely innervates the OFC and modulates its behavioral functions [126,127]. OFC neurons seem to react to specific reward features, such as DRN stimulations of specific frequencies and durations. Furthermore, the addition of DRN stimulation modulates the neuronal responses to natural rewards. These results support the concept that DRN activation produces reward signals that can effectively organize and modulate reward processing in the OFC. Several experiments suggest that OFC neurons distinguished the identities of the different reward signals. The responses of some OFC neurons were tuned to the outcomes of DRN stimulations at different intensities. Moreover, some of the neurons respond selectively to the outcomes of either the artificial reward of DRN stimulation or natural rewards.

Recordings provide evidence that DRN activation produces reinforcement signals that condition, modulate prospective responses to reward outcomes in the OFC, and are suggestive of the importance of DRN neurons in reward processing. In addition to the OFC, the DRN also sends

extensive axonal terminals to many brain areas associated with reward processing, such as the ventral tegmental area (VTA), the nucleus accumbens, and the lateral hypothalamus [128]. Reward signals from the DRN might shape the prospective responses of OFC neurons through the direct projection of the DRN to the OFC or through indirect connections via relays such as the VTA and other brain areas that projected to the OFC.

Up to this point, we explored the body of literature available for neural substrates of reward-based learning. The main issues that arises from the analysis are the following:

- Learning occurs according to hierarchical levels going from reflexes satisfying primary needs, to more complex voluntary movements, which could be goal oriented.
- On the top-most level of voluntary actions learning involves the presence of an Actor Critic architecture, which learns on the base of dopamine to link combinations of stimuli with possible motor behaviors. Neural substrates appears to be Cortex/Basal Ganglia for the Actor and Dopaminergic system for the Critic;
- The evaluation occurring trough dopamine response from Dopaminergic system could be triggered from innate rewards or punishments located in dedicated structures (Later Hypothalamus for rewards and Lateral Habenula for punishments), yet it could come from initial neutral stimuli which acquire the ability to trigger dopamine after Pavlovian conditioning with

innate rewards. Amygdala and Ventral Striatum collaborate to learning of rewards.

- The neural substrate for motivating operations, variables which alter the effectiveness of rewards and punishments seems to be Dorsal Raphè Nucleus for innate and Orbitofrontal Cortex mostly for learned values.

Now that we investigated these structures we want to build a *neuro-computational model* which could validate these issues Since the architecture derived from literature appears deep and complex we are going to adopt some *working hypothesis* leading us to a much simpler model. These hypothesis are not going to denaturalize the core structure of the model. In fact, in the last section which is the *original* contribution of this work we are going to show a *methodology* of experiments in order to evaluate the learning of humans.

We will find out that the results obtained with humans under such methodology could be comparable with those provided by with the computational model Hence the working hypothesis we adopted for simplifying the neuro-computational model will prove to be plausible.

Chapter 3

Computational model

In this chapter, it will be proposed a computational model of the neural architecture we have presented in the previous chapter. Some crucial assumptions will be developed. In performing validation of the model, we will consider two scenarios: the first will show how a system could learn actions from an innate reward system; the second one how neutral stimuli can acquire a value, thus driving the learning of following actions.

3.1. Working hypothesis of the computational model

In the previous chapter, we discussed how an architectural model for reinforcement learning could be layered according to hierarchical levels: now we will make some simplifying assumptions we made in order to obtain the computational model used for the validation.

Hypothesis n.1:

Since we are evaluating the behavior of reward based simple voluntary movements, for which any kind of smooth correction provided by cerebellum is not strictly necessary, we will concentrate on level 3, thus excluding the cerebellar level;

Hypothesis n.2:

The learning task we are accounting for the validation of the model is independent of the context in which it is performed. Therefore, we will not take into account motivating operation phenomena linked to the Orbitofrontal and Prefrontal Cortices We will concentrate on The Posterior and Motor Cortices.

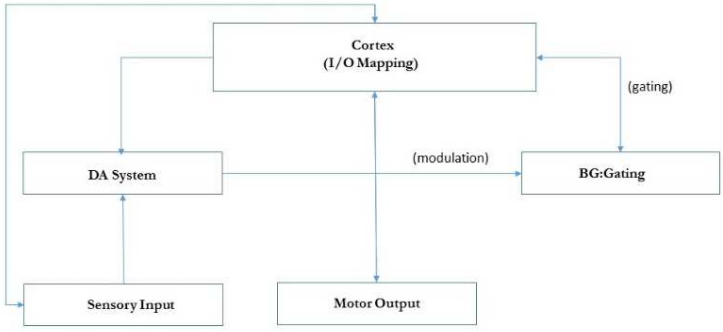


Figure 3.1. A simplified version of the model including level 1 and 3 of the hierarchical system proposed in chapter 2.

Hypothesis n.3:

Since we want to concentrate more on the aspect of reinforcement learning between a set of stimuli and a set of possible voluntary actions, we are not dealing with representing how stimuli from the sensory input level become hierarchically higher perception representations, nor how motor outputs contribute to create complex motor plans. In these conditions we directly have high level representations of sensory inputs (perceptions) and of motor outputs (voluntary movements), which both occur at level of cortex

Hypothesis n.4:

The last hypothesis is that Basal Ganglia could be included in the cortex layer, through a gating rule that selects a voluntary movement for each trial. Therefore, after a reward or a punishment occurs, it will be strengthened only that specific synapse.

Figure 3.2 represents a synthesis of the model after the final hypothesis.

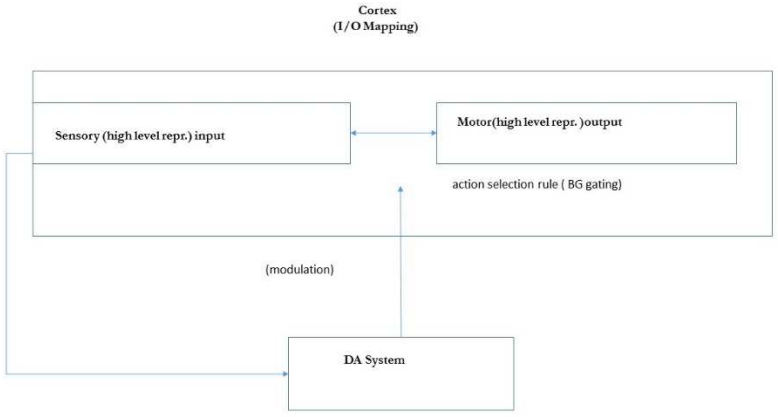


Figure 3.2 (previous page) In our final representation of the model, learning associates voluntary movements in the motor cortex with stimuli combinations in the posterior cortex, after dopamine firing from DA system. BG selection is implemented through a rule on the output.

Now we will proceed with the validations of the model. We will consider two cases:

- Reinforcement learning with innate external rewards/punishments;
- Reinforcement learning with learned rewards/punishments.

3.2 Simulation with innate rewards

In the first validation, we wanted to prove how external rewards and punishments could shape the learning of voluntary movements. The task is quite simple: the system has to learn an action given a sensory stimuli combination in input. By taking in consideration the assumptions made in the previous section, the DA system is made up of VTA/SNc block for producing the delta rule and primary rewards and punishment blocks (Later Hypothalamus and Lateral Habenula). Learning occurs at Input-Output (Cortex) synapses on the base of delta produced by the VTA. The weights are initially set to random values.

In each trial a random stimuli input is selected out of a permutation among the number of possible input stimuli combinations. The output neurons activate according to inputs and initial weighs values. Here the basal ganglia rule occurs; given the output activation values:

1. Normalize these values over the sum of output activities so they sum up to 1;
2. Rearrange these values creating a cumulative distribution;
3. Choose the first value of the distribution, which is greater of a number randomly generated between 0 and 1.

Rewards and punishments are arranged in a matrix (the reward function mentioned in chapter 1) which describes for

a certain stimulus/action pairing if a reward (represented as 1) or a punishment (represented as -1) is available.

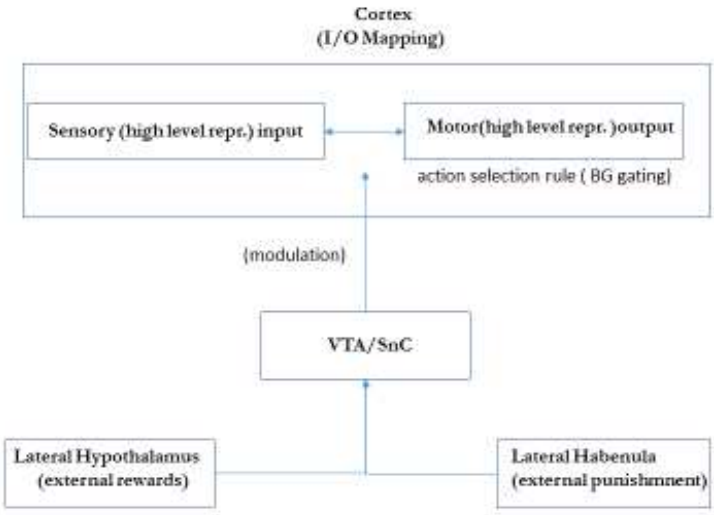


Figure 3.3. Architecture of the model for learning by external rewards and punishments.

3.3 Simulation with learned rewards

In a different simulation scenario, we proved how a set of learned rewards could lead to learn a task. As described in the previous chapters, a way to evaluate how a neutral state could acquire the ability to elicit a response is by learning a sequence

of actions. In this new scenario, the pattern of states presents two steps:

- On odd trials, the system starts from a state (states 1 or 2). No reward nor punishment is provided after the motor output is chosen; the action performed instead leads to a subset of states (states 3 or 4), depending from the action ;
- On even trials the system is in the states (3 or 4) and an action performed could lead to reward or a punishment;

In this case, we added two blocks in the DA systems:

- The Amygdala, which learns in presence of an input state to produce a dopamine burst in the VTA, thus leading to the learning of input output associations.
- The Striatum, which learns to produce in presence of an external reward (punishment) an expectation , in order to cancel out VTA dopamine firing triggered by external reward (punishment)

Therefore, the learning of the system involves synaptic weights between input- output, input- amygdala and input-striatum.

As we reviewed in chapters 1 and 2, in order to learn the CS should have a sustained firing, while also US occurs. In the simulation, since each sequence lasts only two time steps (trials) some simplifications should be made:

- On even trials (at the end of the sequence) the input stimuli (3 and 4) occurs *simultaneously* with the

external reward (punishment), if present. A dopamine burst occurs in the VTA, but in later stages as the striatum learns to produce an expectation, the dopamine burst tends to stay at baseline. Meanwhile amygdala weights tend to increase.

- On odd trials (at the start of the sequence) the input stimuli (1 or 2) is presented simultaneously with the next state stimuli (3 or 4, according to the action), which, if on even trials acquired the ability to activate the Amygdala, it will be able to produce a dopamine burst. The phenomenon of reward prediction (CS producing dopamine burst before US) takes place in real continuous time, but in our simulation time is discrete and a sequence is made up of only two time steps. We need a way to take in account this simultaneous occurrence. Therefore, we are posing that the dopamine burst is shifted back in time from time step 2 to time step 1.
- In chapter 2, we mentioned how Striatal prediction activity seems to occur primarily due to a representation of timing coming from prefrontal Cortex. This representation maps the time elapse between the CS and the US, going to zero when US occurs. Here will make the assumption that striatum is connected to input and that learning in striatum occurs only if reward/punishment is present (on even trials, so keeping the paradigm that Striatum activation predicts US firing and is specific to CS) .

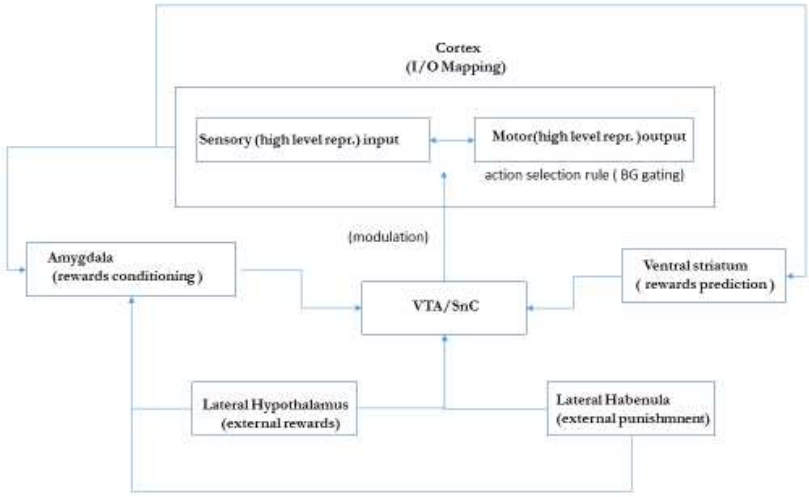


Figure 3.4. Model architecture for conditioned rewards learning paradigm (sequences of actions). The inclusion of amygdala makes possible for sensory inputs to become conditioned rewards, while striatum makes possible for a stimulus to predict the occurrence of an external reward.

3.4 Implementation of the Model and results

The implemented architecture makes use of formalism derived from Emergent software [128], which is able to model neural structures and connections according to physiologically plausible parameters and equations. Using structures and parameters inherited from Emergent the model was implemented in Matlab, so allowing a more versatile yet

limited programming and the possibility to build over the model to interface it with other applications.
The model is structured according to an object-oriented paradigm, in which the neural structures take shape as *layers*, made up by one or more *units*, all layers creating a *network.*

Each simulation consists in a group of epochs, each epoch corresponding to the presentation of a sequence, so that there are two trials for epoch. According to Emergent formalism, each trial is structured in a *minus* and a *plus* phase. In minus phase the net evolves according to the inputs, while in minus the net is clamped to eventual rewards.

For each simulation, we evaluated the goodness of learning according to the learning curve built as the cumulative scoring of rewards (minus punishments) and the dynamics of the weights of input-output synapses and of input-amygdala synapses.

3.4.1 Task with external rewards

In the first task, we evaluated the learning of a set of two possible actions trough innate reinforcement .We evaluated the task varying the cardinality of the set of stimuli: 2 stimuli, 4 stimuli and 8 stimuli.

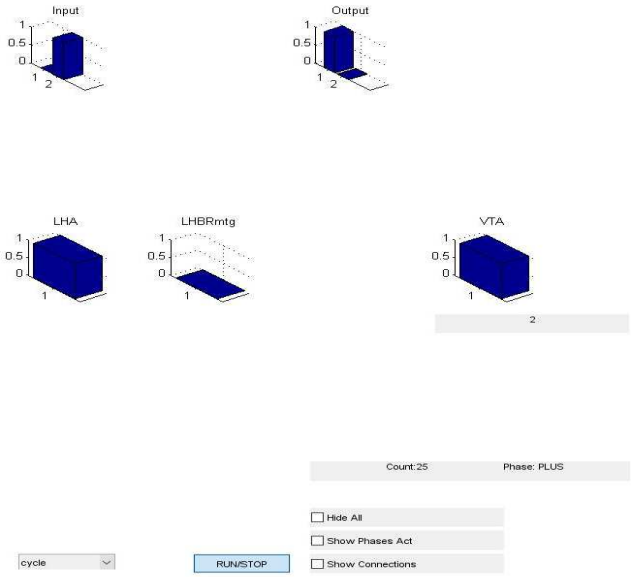


Figure 3.6. Model of learning for innate rewards/punishments. Delta rule produced each trial strengthens/weakens connections between input and output

The innate rewards and punishments are provided by Lateral Hypothalamus and Lateral Habenula blocks, during *plus* phase, while the VTA block produces the delta rule according to the equation:

delta=reward;

where the learning rule is:

*dIj= IN_OUT*delta*net.layers{inIdx}.units(pat).act;* (1)

net.layers {inIdx}.units (pat).act are the input activations with *pat=1: n*

and *IN_OUT* is the learning rate, which is set to 0.1.

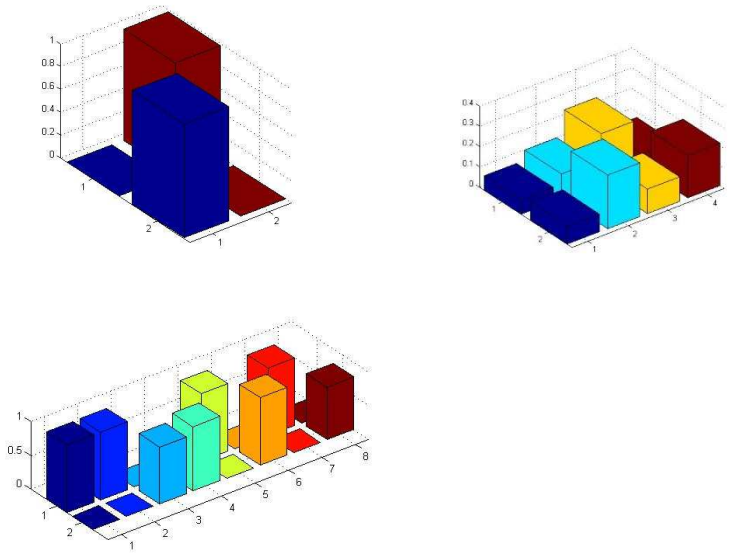


Figure 3.7: input-output weights for external rewards/punishments learning tasks. Results for innate reward simulation: 2 stimuli, 4 stimuli and 8 stimuli. Weights are randomly set.

The rewards are arranged in an nx2 matrix, in which 1 and -1 are randomly generated each execution and equally distributed each row and n is the number of stimuli to learn.

Results were averaged over ten different executions, providing the score showed in the figures below.

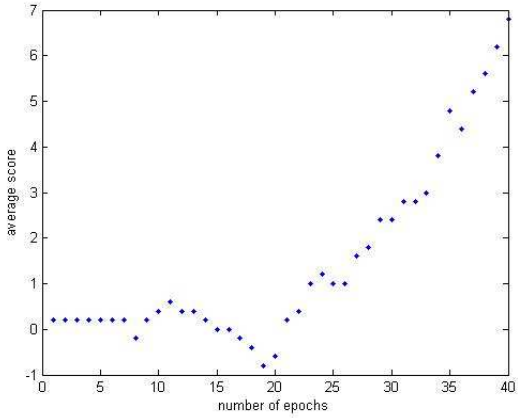


Figure 3.8. Averaged score (on 10 executions) for learning a task with innate reward/punishments. Number of stimuli to learn =2

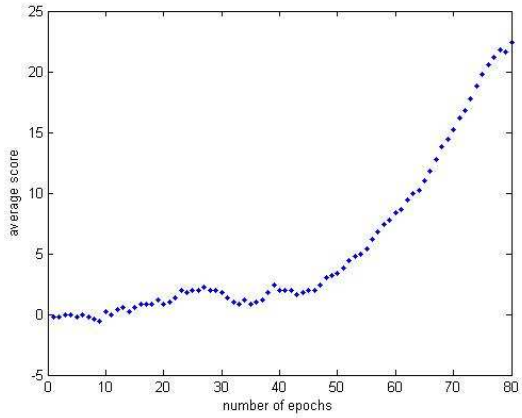


Figure 3.9. Averaged score (on 10 executions) for learning a task with innate reward/punishments. Number of stimuli to learn =4

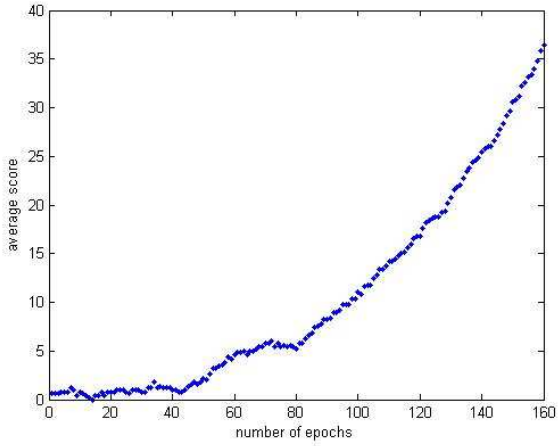


Figure 3.10. Averaged score (on 10 executions) for learning a task with innate reward/punishments. Number of stimuli to learn =8

3.4.2 Task with learned rewards

In the simulation of learning by learned rewards, we evaluated a sequence of two states. Once again, the possible actions are two, but this time the external rewards/punishments are fixed as well as the states reachable performing an action. The possible states are four (*A, B, C, D*) and while two of these states are directly reachable (*A, B*) the other two are reachable only by performing one of the two actions. This info are stored in two matrices:

- a 2x4 matrix for rewards, containing 1 and -1 for innate rewards/ punishments as well as 0 for intermediate rewards;
- a 4x4 matrix for states, which has starting states on rows as well as arriving states on columns and containing as element the index of the action performed (1 or 2) in order to reach such state.

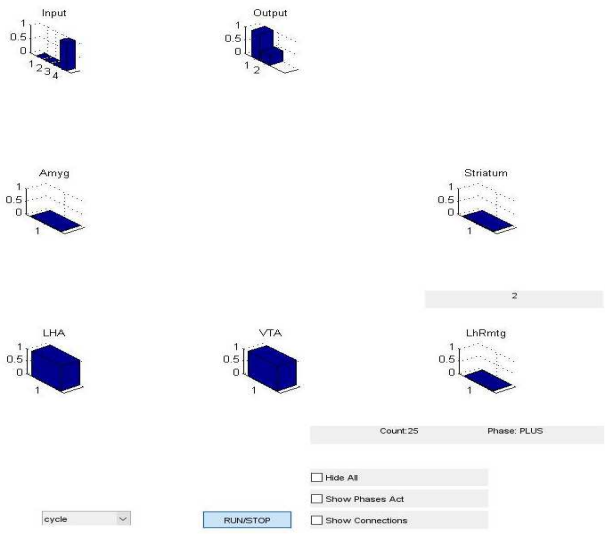


Figure 3.11. Model for learning of learned rewards. Each two trials an innate reward (punishment) is delivered according to the action selected. In odd trials, Amygdala activations drive the delta rule. Striatal activations prevent reward (punishments) from firing dopamine.

The learning rules this time involve input-output, input-amygdala and input-striatum. Delta is calculated each trial and has a different origin depending on the trial: during even trials delta depends on external rewards (punishments) and striatal activations:

delta=deltar-deltas; if reward = =1; (2)

else

delta=deltar+deltas; if rewards = =-1 (3)

where

deltar=net.layers{lhaIdx}.units.act hypothalamus activation
deltar=net.layers{lhbrmtgIdx}.units.act; habenula activation
deltas=net.layers{striaIdx}.units.act; striatum activation

For odd trials, during plus phases the state, which is reached with the action (*C* or *D*), is presented as reward. According to amygdalar weights, the delta now depends on amygdala activations. One important issue is that once a stimulus acquired the ability to elicit a dopamine spike trough amygdala, we have no clue on whether this firing is a punishment or a reward. The strategy implemented is to memorize next state and check the maximum among next state's weights with output units; if the maximum weight leads to a punishment the current stimulus is a learned punishment, otherwise a learned reward:

deltaa =net.layers{amIdx}.units.act; amygdala activation
if rewards (ind, nextstate) ==1
delta=deltaa; (4)
else

delta=-deltaa; (5)
end

where *ind* is the action index and *nextstate* the following state

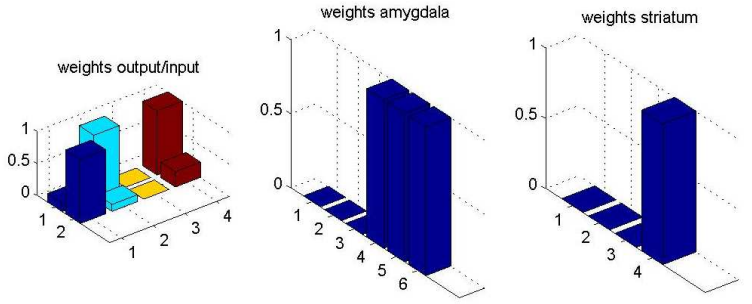
The learning rules are the following:

Learning rule for input-output
*dIj= IN_OUT*delta*net.layers{inIdx}.units(pat).act*; (6)

Learning rules for amigdala and striatum connections to input layer

if reward== 1
*deltastria=IN_STRIA*0.2*net.layers{inIdx}.units(pat).act;* (7)
*deltaam=IN_AM*0.2*net.layers{inIdx}.units(pat).act*; (8)
else
*deltastria=-IN_STRIA*0.2*net.layers{inIdx}.units(pat).act ()* (9)
*deltaam=-IN_AM*0.2*net.layers{inIdx}.units(pat).act; (10)*
end

where *IN_STRIA*, *IN_AM* and *IN_OUT* are the learning rates and are all set to 0.8, a big value in order to speed-up the simulation.



.

Figure 3.12. Weight changes for learning a sequence of 2 stimuli: input-output weights, input-amygdala weights and input-striatum weights.

Amygdala has other two weights connected with hypothalamus (reward conditioning) and habenula (punishment conditioning) which are always set to 1.

The results averaged over 10 executions are shown in the figures below. Especially in the task evaluated, the state and reward matrices are set such that only state *D* leads to a reward through action 1. State *D* could be reached by state *A* trough action 2 and by state *B* trough action 1. So Amygdala weight for *D* is the only one increasing, so input output weights *A*-2, *B*-1 and *D*-1.

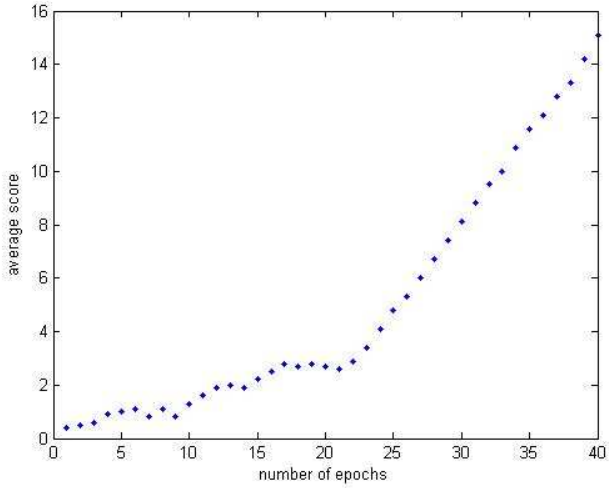


Figure 3.13. Averaged score (on 10 executions) for learning a sequence of 2 stimuli with learned reward/punishments.

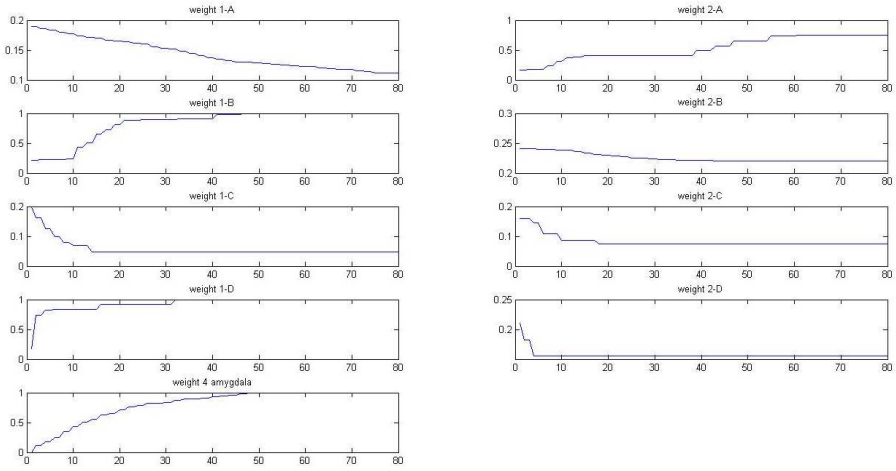


Figure 3.14. Weights between input-output (stimuli *A B C D,* actions 1 and 2) and input amygdala (for stimulus *D*) averaged on 10 subjects

The results of the simulations prove that under the hypothesis laid in section 3.2 the proposed neuro-computational:

- - learns the expected behavior from both innate reward/punishment;
- - learns new values by pairing initially neutral stimuli with innate rewards/punishments;
- .learns the expected behavior from the learned rewards/punishments.

..

In the following chapter we will describe the experiments we have designed and performed for validate the model and to

prove that the model learning dynamics is similar to the one exhibited by human subjects.

Chapter 4

Experiments on learning rate

In the previous chapter, we proved that learning occurs thanks to a system of innate and acquired values endowed with rewarding or punishing properties.

Now we want to go further and prove that this learning paradigm applies to humans as well.

In order to achieve this result experiments were performed on human subjects, evaluating human learning. The aim of these experiments was to evaluate learning curves on human subjects and to compare their learning rates with those exhibited by the model.

Making the comparison between humans and computational model consistent was a pivotal aspect. In chapter 3, while building our computational model (figure 3.2), some assumptions were made on the architectural model of chapter 2 (figure 2.1), especially for the Cortex. Since modeling Cortex would require a considerable amount of complexity, due to its deep hierarchical layered structure both for input sensory perception that for output behaviors, we opted for a simple one-layered neural network with stimuli

combinations in input and only possible actions on output. The network was fully connected, and in the first stages of learning all the weights had the same probability; as reward/punishment occurred, the weights changed as well. Now the point is, does this simplified input-output scheme could apply also for modeling human subjects during the experiment? The answer is 'yes', but we need to carefully design the experiment, satisfying the following requirements:

- Making the subject choose among a limited set of possible actions;
- Reducing the set of stimuli provided to the subject ;
- Unbundling the task from every possible type of semantic context which could recall any previous skill or knowledge of the subject;

Under these requirements, the experiment involved a learning task in which groups of visual stimuli were presented to the subjects, which had to press a button on a device. The goal of the task was to learn to associate those stimuli with the right button: being limited in the set of stimuli and having the same output motors, the subjects were so in the same conditions as the computer. The choice of visual stimuli was performed in a way to make their semantic as simple as possible. In this way the perception, which in human brain is made up by a quite complex hierarchy of neural layers, should collapse to a single layer making the results comparable. The human brain easily recognizes edges, so an effort was put toward shaping the visual stimuli as geometrically simple as possible, experiment after experiment. Figure 4.1 shows the scheme of the experiment.

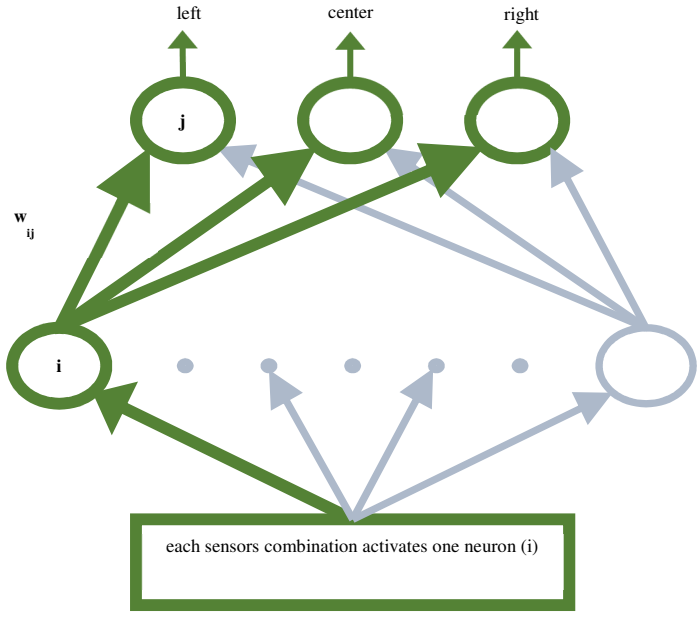


Figure 4.1 The experimental setting model: the subject has to press a button (in the figure, one out of 3) after being presented with a visual stimulus, which is an easily semantic understandable perception (combination of sensors). The task consists in learning to match each stimuli with the right button

We will further go in detail in the following sections and later discuss the results providing important findings.

4.1 Experimental settings

In the following sections, information will be provided about the different experimental tasks and settings.

4.1.1 First experiment: Robotic Head

The first group of experiments involved the use of a robotic head placed in a box. The head was equipped with three light sensors and could rotate around a vertical axis via a servomotor. Two light sensors were placed like two “eyes” on both sides of the head, while the third sensor was placed on the midline like a “mouth”. Three lights were placed in front and on both sides of the head respectively. The participants, being unaware of the presence of the head in the box, only had a visual cue, in form of three vertical bars on the monitor, representing respectively the light intensity of 2 “eyes” photo-sensors and the head position respectively. In total, there were 14 combinations of stimuli to learn. On each trial, the computer randomly turned on one of the three lights. The user had to press one out of 3 keyboard keys (*A, S, D*), turning the robotic head in one of the 3 possible positions : left, center and right. Whenever the head turned in the direction of the light, the user got a positive feedback, in the form of a visual and auditory stimulus.

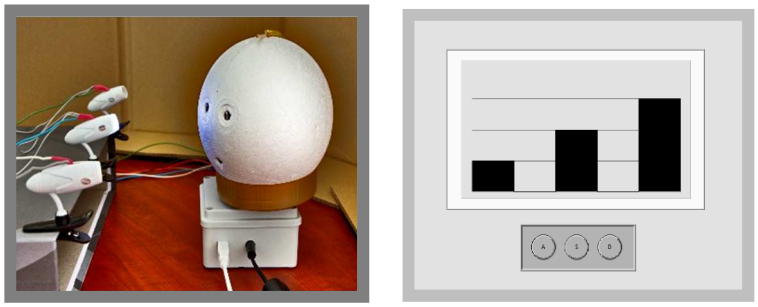


Figure 4.2. The first experiment involved turning a robotic head toward lights; the scene was hidden in a box to the user, which had only visual cues on monitors appearing as histograms (stimuli to learn) and buttons (action to select).

The visual feedback occurred in form of a red (negative) or green (positive) frame, while the auditory feedback as a buzz sound.

Each session lasted 20 minutes, and a total of 40 participants was tested. Software interface was implemented with *LabVIEW*[139] .

4.1.2 Second experiment: two buttons task

The second group of experiments was software only (robot head not involved) and this time users had to choose among a pair of two buttons (A, D).

Figure 4.3 In the second experiment a software only scenario was implemented, involving the pressure of one out of 2 buttons, for 2, 4 and 8 stimuli. The 8 different stimuli are showed in the figure.

The software used was the same of the previous experiment, as well as the interface. The task consisted into learning to associate sequences of stimuli to with the two buttons, in three increasing difficulty scenarios: first 2, than 4 and finally 8 stimuli. Each task lasted 10 minutes for each set of stimuli. So 30 minutes overall occurred to conclude the experiment.

4.1.3 Third experiment: cloche with geometric figures

Finally, a third group of experiments included using a cloche with two buttons and responding to three block of visual stimuli, once again according to three cases: 2, 4 and 8 stimuli. This time the experiment was not time limited but ended after the users had given a fixed number of responses .These number varied according to the number of stimuli of the task. Each presentation of a stimulus was a trial, and a group of presentations of all the set of stimuli (2, 4 or 8 stimuli) was called epoch. So each epoch was a permutation over the set of stimuli(trials), for a total of 20 epochs each execution; so 40 stimuli presentations, trials, occurred for the 2 stimuli learning task case, 80 for the 4 stimuli learning task and 160 for the 8 stimuli learning task the. Two scenarios were evaluated: in the first one, the associations between buttons and stimuli were the same for all the executions (fixed), while in the second they changed each time (random).

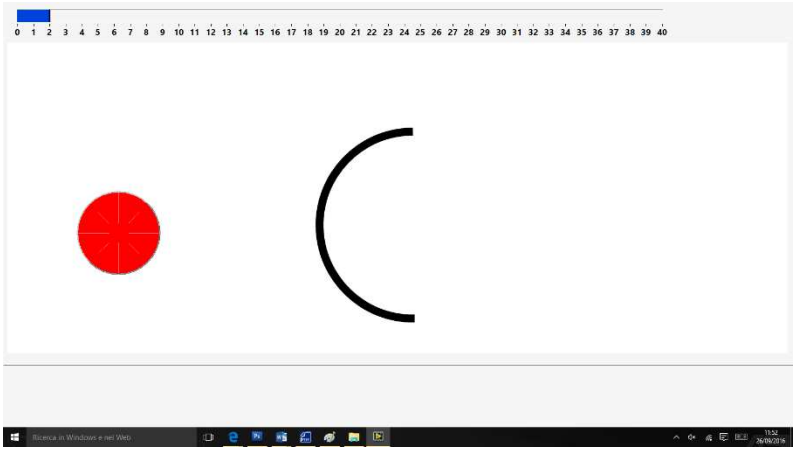


Figure 4.4 Computer interface for last experiment: a geometric figure is shown on monitor each trial and the user has to press a button out of two on a cloche. Feedback is provided through a circle at the side of the pressed button on the GUI (green for reward, red for punishment.)

Figures were chosen in order to have a semantic as simple as possible, yet to not have any kind of orientation (for example pointing left or right) such to associate them to a certain button. Different sets of figures were used for the 2, 4 and 8 stimuli tasks.

For the three tasks reward consisted once again in a sound (‘yeah’ for reward or a noise for punishment), and a visual feedback, (a circle appearing on the side of the pressed button, green for reward and red for punishment).

Data from the experiments was then fitted in order to obtain learning curves of the task, learning to associate stimuli with actions. We will now discuss these results.

4.2 Fitting of data

A fitting on data was performed in order to obtain the learning curves. According to [129] and [130], we adopted as fitting function the exponential one, mostly due to its nonlinear form that can describe at best a great range of tasks.

4.2.1 Base model: fitting for 2 stimuli

The function evaluated for the fitting of the experiments is the following:

$$w_t = m+(w_0 - m)\, k^t \qquad (1)$$

where:

m is the maximum value for the weight (set to 1);
w_0 is the initial weight;
t is the trial ;
$(1 - k)$ is the learning rate to estimate

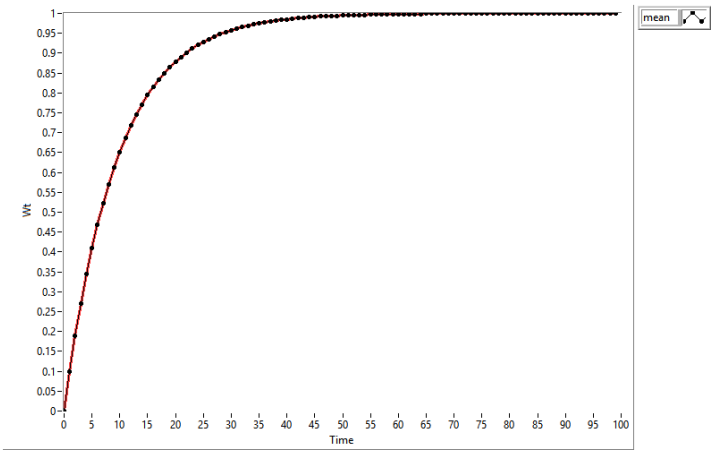


Figure 4.5. Fitting with an exponential function for the 2 stimuli task. As the increase of an input-output weight is proportional to the sum of rewards, it could represent somehow the goodness of the learning.

As the equations shows, we took in consideration the input-output weight equation (1) of Chapter 3, transforming the iterative form in an explicit function.We considered the weight of a synapse responsible for correct answers as the probability of giving the correct answer.

4.2.2 Model with n stimuli to learn

The learning of a specific stimulus follows the previous trend but updating occurs every n presentations because there are n interwoven stimuli. The assumption is that if there are multiple stimuli to learn, we can expect that the weight of an input output synapse will be strengthened averagely after n rewards

$$w_t = m + (w_0 - m)\, k^{t/n} \qquad (2)$$

where the meaning of the symbols is the same as in eq (1), n is the number of different stimuli and $(1 - k^{1/n})$ is the learning rate to estimate.

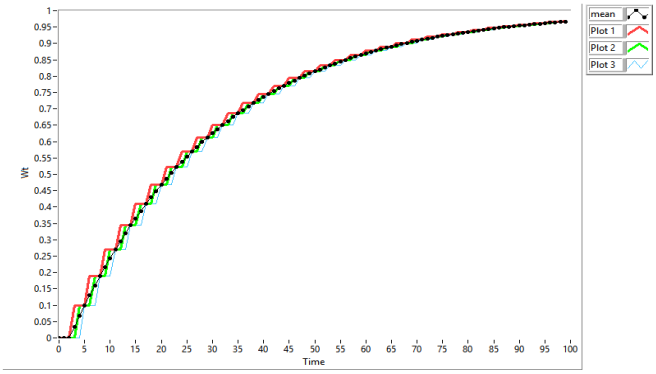


Figure 4.6 The weight increase in the case of 3 stimuli to learn. The dotted curve is the average of the 3 curves.

4.2.3 Hyperbolic Fitting

A second approach was to consider that the probability of giving the correct answer was the ratio between the weight of the synapse responsible for the correct answer and the sum of the weights of all the synapses activated by a stimulus. This approach lead to fit the experimental data with the following function:

$$w\ (t+1) = \frac{m+(w(0)-m)\ (1-a)^t}{m+(w(0)-m)\ (1-a)^t+(n-1)w(0)}$$

where
w (0) = initial, minimum value for *w (t)*;
m = maximum possible value for *w (t)* (set to 1);
n = number of possible actions
t = number of correct responses;
a = parameter to estimate for best fitting.

4.3 Results

For each experimental condition (2-stimuli learning, 4-stimuli learning, 8-stimuli learning, 14-stimuli/robotic head), we

obtained an array of data, containing the averages among all participants of the first responses following each reward; this array represented an estimate of the probability of giving the correct answer updated after each reward. Such solution was adopted in order to put data for all participants in the same conditions: since in the first experiments time was fixed, the number of responses given (and reward obtained) could vary according to how 'trigger happy' some participant could be.

After normalization, only responses following rewards are taken into account, so the problem is no more, since subjects are evaluated only on the number of rewards purchased.

These arrays were then fitted with the two different approaches.

4.3.1 Results for robotic and software only experiments

For the first experiment, robotic head turning task, a group of 41 subject was evaluated.

For the second experiment, software only, a group of 49 subjects was evaluated.

The results provided with curves fitted with exponential and hyperbolic fitting are showed in the figures below.

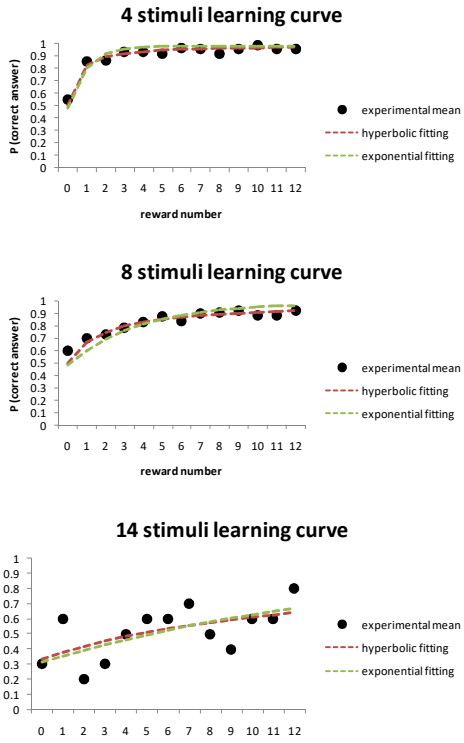


Figure 4.7. Results for the fitting in the first experiment (14 stimuli) and in the 4 and 8 cases for second experiment. Dots represents averaged probabilities of giving a correct answer. The dashed figures represent the exponential (green) and hyperbolic (red) fittings.

The results show that for 4 and 8 stimuli both the fitting proved good, while in the case of 14 stimuli (robot head turning task) both curves were not able to approximate correctly data. This is also because, while in the first two cases the probability of giving a correct response quickly converges

to a maximum (after nearly 2 rewards in the 4-stimuli case and 4 rewards in the 8-stimuli case), in the last case it never stabilizes, and oscillates around an average value of approximately 0.6.

4.3.2 Results for experiments with cloche

In the experiment with cloche, we evaluated results for both the random and the fixed presentations of stimuli. To avoid attention shift due to the perceived (by the subject) complexity of the task which we have observed during the experiment, we considered that the task was learned by a subject if the participant is able to consistently give the correct answer for a certain number of epochs, which depends on the task complexity. In the experiments reported below, we have assumed for such a period the following:

- 5 consecutive epochs, in the 2-stimuli learning task ;
- 4 consecutive epochs, in the 4-stimuli learning task ;
- 3 consecutive epochs, in the 8-stimuli learning task;

In the figures are showed the curves for the three tasks in the fixed and random scenarios, for one execution.

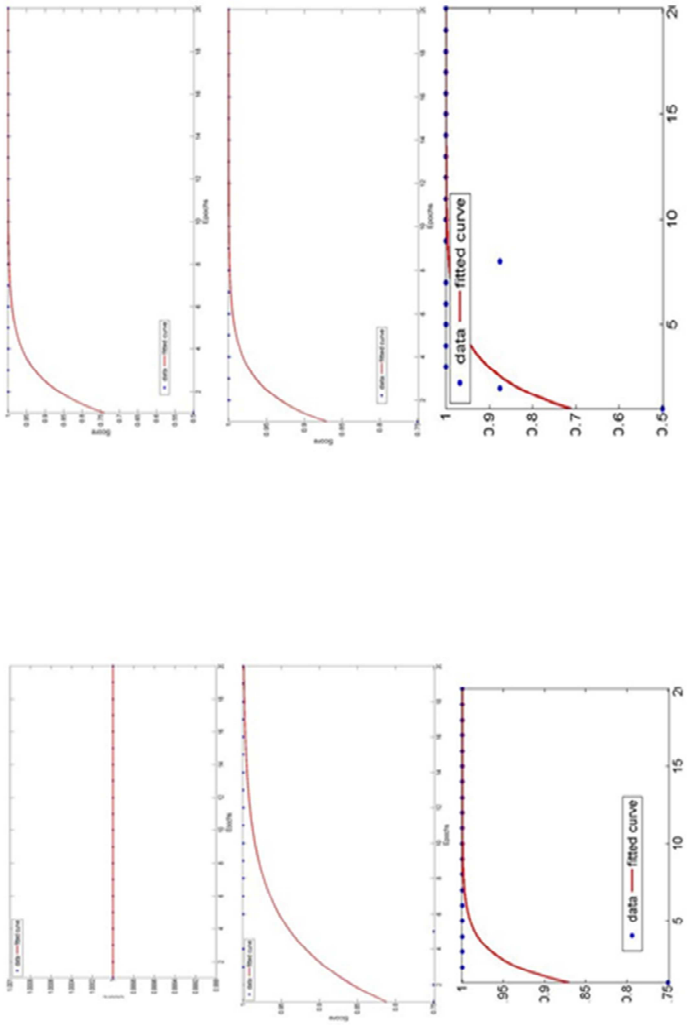


Figure 4.8. Learning curves with exponential fitting for one execution: left column shows results for 2, 4, and 8-stimuli learning tasks (fixed), right column shows results for fixed scenario

4.4 Discussion

Figures 4.8 and 4.9 show the estimation for learning rates. Here, we confronted averaged data with theoretical value coming from the equation:

$$(1 - k^{1/n}) \qquad (3)$$

where *k= average [(1 -average learning rate in the n stimuli test)* ^{n}*]*

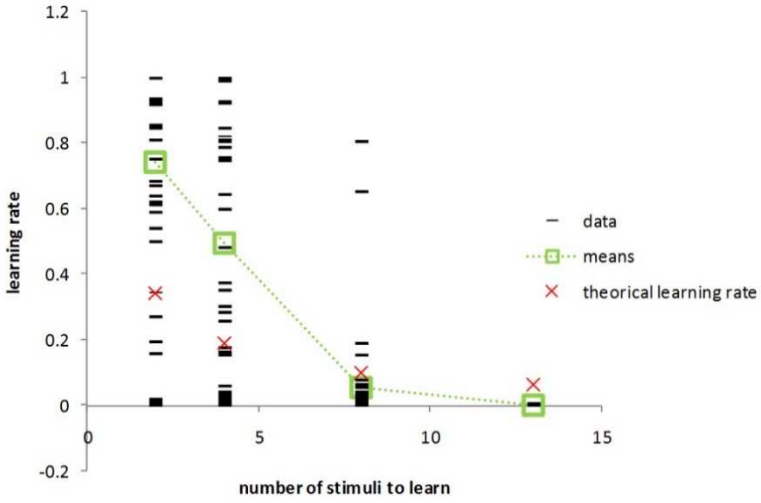


Figure 4.9. Estimated learning rates for the task of robotic head turning and software (2, 4, 8 stimuli).

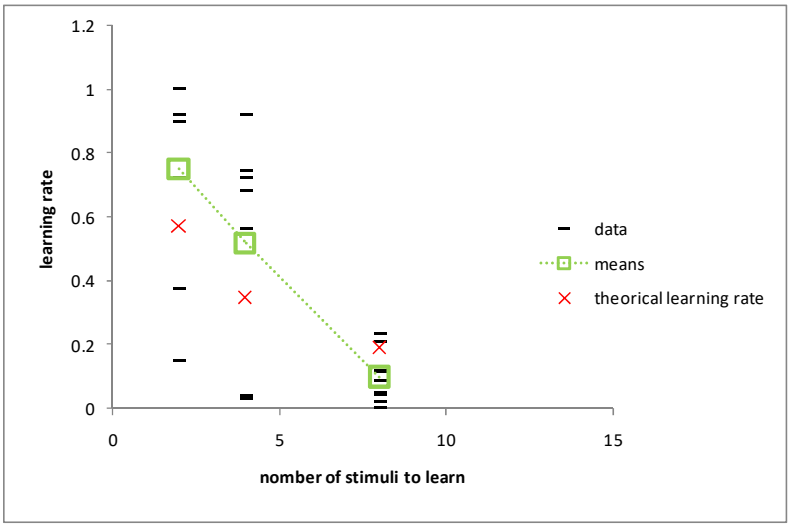


Figure 4.10. Estimated learning rates for the cloche task

The results show an important finding: learning rate is not fixed, but varies with the difficulty of the task, precisely with the increasing number of stimuli to learn. The experimental values of learning rate in the software and robotic head turning scenarios are set above the theoretical value for both the 2 and 4 stimuli tasks, but they drop under that values for the 8 and 14 stimuli tasks.

The same applies for the cloche scenario, for which in 2 and 4 stimuli tasks the experimental values are once again above theoretical values, while in the 8 stimuli task they fall below.

There appears to be no particular difference between the fixed and random presentations as we obtained:

- for random presentations the average learning rates appear to be 0.66 for 2 stimuli, 0.34 for 4 stimuli and 0.22 for 8 stimuli
- for fixed presentations, the average learning rates appear to be 0.7 for 2 stimuli, 0.41 for 4 stimuli and 0.32 for 8 stimuli.

The results found are not so surprising, in fact in literature a body of work [132-134] shows how human learning abilities strictly depends on how much information we are able to process simultaneously and on how much information we already have memorized.

Especially it looks like that a transmission containing more than 7 elements cannot be entirely learned (Probability of correct answer<0.05). This is compatible with the fact that learning curves decrease when the number of stimuli to learn is higher than 7 (8 and 14 stimuli in our experiments). These finding seems to suggest that the role of *working memory*, a neural function that seems to be implemented mostly by prefrontal cortex [135-138], may be fundamental for keeping representations of stimuli for the time needed for learning.

Conclusions

In this thesis, we raised several questions regarding where and how value information can drive human learning.

We found out that learning could occur through the positive or negative feedback effect of specific stimuli, called rewards and punishments, and that they could be delivered as the consequence of a specific action or independently of the meaning of any action, in an associative mechanism.

From the analysis of literature, we discovered that these stimuli could come from innate needs, yet they could be initially neutral stimuli that acquired the ability to produce a reward or a punishment after a Pavlovian conditioning. Usually these acquired reward/punishments occur as the intermediate rewards in a sequence of actions, at the end of which the innate reward is delivered. In case the innate reward becomes less and less available, the acquired reward begins to lose its reinforcing abilities (the same applies for acquired punishments). This phenomenon is called extinction and is strictly linked to the context in which the action takes place, called motivating operation. We also found out that motivating operations could be innate or acquired.

We then explored the neural substrates of reinforcement learning and found out that dopamine acts as the rewarding (burst) or punishing (dip) signal. A specific system is involved in the firing of dopamine and involves the VTA and the

Substantia Nigra Pars Compacta, while the Amygdala is responsible for the learning of rewards.

Dopamine then stimulates the connection between combinations of stimuli in the Posterior Cortex and possible motor actions in the Motor Cortex. At beginning of learning, all the matchings have the same probabilities, then some become more frequent as reward and punishments are delivered. The Basal Ganglia are responsible for the selection of some actions in favor of others. We also hypothesized that voluntary movements are part of a hierarchical system that comprises from elementary reflexes to goal directed behaviors, and we gave a glimpse of the neural structures involved in the other levels.

We tested our hypothesis about the neural regions implementing these systems by simulations. We built a simple task, where certain stimuli states could be reached performing actions. The first task was on innate rewards, while the second simulation proved that learned reward could stimulate to learn a sequence.

These simulations were performed by means of a neural computational model that was built starting from literature findings and applying some simplifications that did not alter the overall meaning of a reward learning system.

Finally, we wanted to evaluate the extent to which the performance of the model compared with that of human subjects. For this purpose, we performed a set of experiments with human subjects. The experiments were carefully designed to bring humans and computer simulations in the same conditions, collapsing the complex hierarchy of layers for both perception and motor planning.

The results of our experiments showed that the human learning rate is not constant, as hypothesized in the computer simulations. In particular, it appears that by increasing the number of stimuli to learn simultaneously, humans need a greater number of presentations of each stimulus to reach the maximum learning. This finding provides further support about the working memory, a short time type of memory specific for the learning of movements, located in the prefrontal cortex. Previous researches found out that when humans have to learn to memorize simultaneous stimuli, the limit is around number 7, after which the performance falls down. Our findings confirm that findings in a different experimental setting.

As we did not take into account the working memory and its correlates, a further step in this direction could be to implement a model including some kind of memory properties.

In the current model we made the simplifying assumptions that perception is a combinations of stimuli: it could be interesting to explore these bottom up phenomenon of how simple stimuli could be recognized in higher order categories. Actually in the primate brain, visual information in the cortex flows through a *cortical hierarchy*[140]. These areas include V2, V3, V4 and area V5/MT (the exact connectivity depends on the species of the animal). These secondary visual areas (collectively termed the *extrastriate visual cortex*) process a wide variety of *visual primitives*. Neurons in V1 and V2 respond selectively to bars of specific orientations, or combinations of bars. These are believed to support edge and corner detection. Similarly, basic information about color and motion is processed here. As visual information passes

forward through the visual hierarchy, the complexity of the neural representations increases. Whereas a V1 neuron may respond selectively to a line segment of a particular orientation in a particular retinotopic location, neurons in the lateral occipital complex respond selectively to complete object (e.g., a figure drawing), and neurons in visual association cortex may respond selectively to human faces, or to a particular object .

DNN(Deep Neural Networks) [141],are computer vision models in which model neuron tuning properties are set by supervised learning without manual intervention. DNNs are the best performing models on computer vision object recognition benchmarks and yield human performance levels on object categorization Therefore a visual computational model based on DNNs would be a good option for a proper characterization of human perception.

Bibliography

[1] Levitis, Daniel A., William Z. Lidicker, and Glenn Freund. "Behavioral biologists do not agree on what constitutes behavior." Animal behavior 78.1 (2009): 103-110.

[2] Cooper J.O., Heron T.E., Heward W.L. (2007). Applied Behavior Analysis (2nd ed.). Prentice Hall. ISBN 0-13-142113-1.

[3] Kandel E, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ. Principles of Neural Science, Fifth edition, chapter 35 “Spinal Reflexes”. McGraw-Hill, New York, NY, USA, 2012.

[4]Ropper AH, Brown RH. Principles of Neurology, Eight edition, chapter 28 “Normal development of the nervous system -Development during the neonatal period, infancy, and early childhood”. McGraw-Hill, New York, NY, USA, 2005.

[5]Thompson AK, Wolpaw JR. Operant conditioning of spinal reflexes: from basic science to clinical therapy. Front Integr Neurosci. 2014; 8: 1-25.

[6]Pavlov PI (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. Ann Neurosci. 2010 Jul;17(3): 136-41.

[7] Skinner BF. The Behavior of Organisms: An Experimental Analysis. Appleton-Century-Crofts, New York, NY, USA, 1938

[8]Michael J. Distinguishing between discriminative and motivational functions of stimuli. Journal of the Experimental Analysis of Behavior. 1982;37:149–155.

[9] Sundberg M. L. The application of establishing operations. The Behavior Analyst. 1993;16:211–214.

[10]Hart B, Risley T. R. Incidental teaching of language in preschool. Journal of Applied Behavior Analysis. 1975;8:411–420

[11]Verschure, Paul FMJ, Cyriel MA Pennartz, and Giovanni Pezzulo. "The why, what, where, when and how of goal-directed choice: neuronal and computational principles." Phil. Trans. R. Soc. B 369.1655 (2014): 20130483.

[12]Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry. MIT AI Laboratory.

[13]Gurney, K. N., Prescott, T. J., Wickens, J. R. and Redgrave, P. (2004). Computational models of the basal ganglia: from robots to membranes. Trends Neurosci 27(8): 453-459.

[14]Bellman, R. E. (1957). Dynamic Programming. Princeton University Press, Princeton, NJ

[15]Balkenius, C. and Moren, J. (1998). Computational models of classical conditioning: A comparative study. LUCS 62 ISSN 1101-8453, Lund University Cognitive Studies.

[16]Klopf, A. H. (1972). Brain function and adaptive systems - a heterostatic theory. Technical report, Air Force Cambridge Research Laboratories Special Report No. 133, Defense Technical Information Center, Cameron Station, Alexandria, VA 22304.

[17]Klopf, A. H. (1982). The hedonistic neuron: A theory of memory, learning, and intelligence. Hemisphere, Washington, DC.

[18]Klopf, A. H. (1986). A drive-reinforcement model of single neuron function. In Denker, J. S., editor, Neural networks for computing: AIP Conf. Proc. , volume 151. New York: American Institute of Physics.

[19]Klopf, A. H. (1988). A neuronal model of classical conditioning. Psychobiol., 16(2):85-123.

[20]Widrow, Bernard, and Marcian E. Hoff. "Adaptive switching circuits." IRE WESCON convention record. Vol. 4. No. 1. 1960.

[21]Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A Survey. Journal of Artificial Intelligence Research, 4:237-285.

[22]Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning. Variations in effectiveness of reinforcement and non-reinforcement. In A. Black & W.F. Prokasky, Jr. (eds.), Classical Conditioning II New York: Appleton-Century-Crofts.

[23]Watkins, C. J. C. H. (1989). Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge, England.

[24]Witten, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. Information and Control, 34:286-295

[25]Barto, Andrew G., Richard S. Sutton, and Charles W. Anderson. "Neuronlike adaptive elements that can solve difficult learning control problems." IEEE transactions on systems, man, and cybernetics 5 (1983): 834-846.

[26]Touzet, C. and Santos, J. F. (2001). Q-learning and Robotics. IJCNN'99, European Simulation Symposium, Marseille.

[27]Schultz, W. (2002). Getting formal with dopamine and reward. Neuron, 36:241-263.

[28] O'Reilly, Randall C., et al. "PVLV: the primary value and learned value Pavlovian learning algorithm." B

[29] Redgrave, P and Gurney, K.N. (2007). What does the short-latency dopamine signal reinforce? Nature Reviews Neuroscience. In press.ehavioral neuroscience 121.1 (2007): 31.

[30] Porr, B. and Wörgötter, F. (2007). Learning with Relevance: Using a third factor to stabilise Hebbian learning. Neural Comp. In press.

[31] Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning: An Introduction. Bradford Books, MIT Press, Cambridge, MA, 2002 edition

[32] Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, 275:213-215.

[33]Hazy, Thomas E., Michael J. Frank, and Randall C. O'Reilly. "Neural mechanisms of acquired phasic dopamine responses in learning." Neuroscience & Biobehavioral Reviews 34.5 (2010): 701-720.

[34]https://grey.colorado.edu/CompCogNeuro/index.php/CCNBook/Sims/Motor/PVLV

[35]Kandel E, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ. Principles of Neural Science, Fifth edition, chapter 35 "Spinal Reflexes". McGraw-Hill, New York, NY, USA, 2012.

[36]Downman, C. B. B., and B. A. McSwiney. "Reflexes elicited by visceral stimulation in the acute spinal animal." The Journal of physiology 105.1 (1946): 80.

[37]Sherrington, Charles S. "Address on the spinal animal." Medico-chirurgical transactions 82 (1899): 449-478.

[38]Thompson, Aiko K., and Jonathan R. Wolpaw. "Operant conditioning of spinal reflexes: from basic science to clinical therapy." Front Integr Neurosci 8 (2014): 25.

[39]Christian, Kimberly M., and Richard F. Thompson. "Neural substrates of eyeblink conditioning: acquisition and retention." Learning & memory 10.6 (2003): 427-455.

[40]Thompson, R. F., and J. E. Steinmetz. "The role of the cerebellum in classical conditioning of discrete behavioral responses." Neuroscience 162.3 (2009): 732-755.

[41]Yeo, Christopher H., and Germund Hesslow. "Cerebellum and conditioned reflexes." Trends in cognitive sciences 2.9 (1998): 322-330.

[42]Burguière, Eric, et al. "Role of the cerebellar cortex in conditioned goal-directed behavior." Journal of Neuroscience 30.40 (2010): 13265-13271.

[43]Ito, Masao. "Cerebellar long-term depression: characterization, signal transduction, and functional roles." Physiological reviews 81.3 (2001): 1143-1195.
[44]Armano, S., et al. "Long-term potentiation of intrinsic excitability at the mossy fiber–granule cell synapse of rat cerebellum." Journal of Neuroscience 20.14 (2000): 5208-5216.
[45]Fujita, M. "Adaptive filter model of the cerebellum." Biological cybernetics 45.3 (1982): 195-206.
[46]Glickstein, Mitchell. "The cerebellum and motor learning." Current opinion in neurobiology 2.6 (1992): 802-806.
[47]Andersson, G., Martin Garwicz, and Germund Hesslow. "Evidence for a GABA-mediated cerebellar inhibition of the inferior olive in the cat." Experimental brain research 72.3 (1988): 450-456.
[48]Herreros, Ivan, and Paul FMJ Verschure. "Nucleo-olivary inhibition balances the interaction between the reactive and adaptive layers in motor control." Neural Networks 47 (2013): 64-71.
[49]Garwicz, Martin, Anders Levinsson, and Jens Schouenborg. "Common principles of sensory encoding in spinal reflex modules and cerebellar climbing fibres." The Journal of physiology 540.3 (2002): 1061-1069.
[50] De Zeeuw, Chris I., et al. "Microcircuitry and function of the inferior olive." Trends in neurosciences 21.9 (1998): 391-400.
[51]Dasgupta, Sakyasingha, Florentin Wörgötter, and Poramate Manoonpong. "Neuromodulatory adaptive combination of correlation-based learning in cerebellum and reward-based learning in basal ganglia for goal-directed behavior control." (2014).

[52]Verduzco-Flores, Sergio O., and Randall C. O'Reilly. "How the credit assignment problems in motor control could be solved after the cerebellum predicts increases in error." Frontiers in computational neuroscience 9 (2015).
[53]Sterling P, Eyer J. 1988 "Allostasis: a new paradigm to explain arousal pathology." Psychiatry Neurosci. 30, 315 – 318.
[54]Sanchez-Fibla M, Bernardet U, Wasserman E, Pelc T, Mintz M, Jackson JC, Lansink C, Pennartz C, Verschure PFJ. 2010" Allostatic control for robot behavior regulation: a comparative rodent-robot study. " Adv.Complex Syst. 13, 377 –403.
[55]Fukuda, M., Ono, T., 1993. " Amygdala-hypothalamic control of feeding behavior in
monkey: single cell responses before and after reversible blockade of temporal
cortex or amygdala projections. " Behavioural Brain Research 55 (2), 141–233.
[56]Norgren, R., 1976. " Taste pathways to hypothalamus and amygdala. " Journal of
Comparative Neurology 166 (1), 17–30
[57]Schultz, W., Romo, R., 1990. "Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. "Journal of Neurophysiology 63 (3), 607–624.
[58]Mora, F., Rolls, E.T., Burton, M.J., 1976. " Modulation during learning of the responses
of neurons in the lateral hypothalamus to the sight of food. " Experimental Neurology 53 (2), 508–519
[59]N Nakamura, K., Ono, T., 1986. " Lateral hypothalamus neuron involvement in integration of natural and artificial rewards and cue signals. " Journal of Neurophysiology 55, 163–181eurology 53 (2), 508–519

[60]Rolls, E.T., Burton, M.J., Mora, F., 1976. " Hypothalamic neuronal responses associated
with the sight of food. " Brain Research 111 (1), 53–66.
[61]Phillipson, O.T., 1978. " Afferent projections to a10 dopaminergic neurons in the rat as shown by the retrograde transport of horseradish peroxidase. " Neuroscience Letters 9, 353–359.
[62]Semba, K., Fibiger, H., 1992. " Afferent connections of the laterodorsal and the pedunculopontine tegmental nuclei in the rat: a retro- and antero-grade transport and immuno - histochemical study. " Journal of Comparative Neurology 323, 387–410

[63]Floresco, S.B., West, A.R., Ash, B., Moore, H., Grace, A.A., 2003. " Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. " Nature Neuroscience 6, 968–973
[64]Stuber, G., Klankder, M., de Ridder, B., Bowers, M.S., Joosten, R.N., Feenstra, M.G.,
[65]Bonci, A., 2008. Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. Science 321, 1690–1692
[66]Matsumoto, Masayuki, and Okihide Hikosaka. "Lateral habenula as a source of negative reward signals in dopamine neurons." Nature 447.7148 (2007): 1111-1115.

[67]Andres, KH; During, MV; Veh, RW (1999). "Subnuclear organization of the rat habenular complexes". Journal of Comparative Neurology. 407 (1): 130–150.

[68]Bourdy R, Barrot M (November 2012). "A new control center for dopaminergic systems: pulling the VTA by the tail". Trends Neurosci. **35** (11): 681–690

[69]Barrot M, Sesack SR, Georges F, Pistis M, Hong S, Jhou TC (October 2012). "Braking dopamine systems: a new GABA master structure for mesolimbic and nigrostriatal functions". J. Neurosci. **32** (41): 14094–14101
[70]Heldt, S.A., Ressler, K.J., 2007. " Lesions of the habenula produce stress- and dopamine-dependent alterations in prepulse inhibition and locomotion. " Brain Research 1073 (4), 229–239
[71]Ullsperger, M; von Cramon, DY (2003). "Error monitoring using external feedback: Specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging". Journal of Neuroscience. 23 (10): 4308–4314.
[72]Shepard, P.D., Holcomb, H.H., Gold, J.M., 2006. " The presence of absence: habenular
regulation of dopamine neurons and the encoding of negative outcomes. " Schizophrenia Bulletin 32 (3), 417–421
[73]Ono, T., Nishijo, H., Uwano, T., 1995. " Amygdala role in conditioned associative learning. " Progress in Neurobiology 46, 401–422.
[74]Amaral, D.G., Price, J.L., Pitkanen, A., Carmichael, S.T., 1992. Anatomical organization
of the primate amygdaloid complex. In: The Amygdala: Neurobiological Aspects
of Emotion, Memory, and Mental Dysfunction. Wiley-Liss, New York, pp. 1–66
[75]Pitkanen, A., 2000. Connectivity of the rat amygdaloid complex. In: Aggleton, J.P.(Ed.), The Amygdala: A Functional Approach. 2nd ed. Oxford University Press,
Oxford, pp. 31–115
[76]Fudge, J.L., Haber, S.N., 2000. The central nucleus of the amygdala projection to dopamine subpopulations in primates. Neuroscience 97, 479–494

[77]Wallace, D.M., Magnuson, D.J., Gray, T.S., 1992. Organization of amygdaloid projections to brainstem dopaminergic, noradrenergic, and adrenergic cell groups in the rat. Brain Research Bulletin 28, 447–454
[78]Petrovich, G.D., Setlow, B., Holland, P.C., Gallagher, M., 2002. Amygdalo-hypothalamic circuit allows learned cues to override satiety and promote eating. Journal of Neuroscience 22 (19), 8748–8753.
[79]Ahn, S., Phillips, A.G., 2003. Independent modulation of basal and feeding-evoked dopamine efflux in the nucleus accumbens and medial prefrontal cortex by the central and basolateral amygdalar nuclei in the rat. Neuroscience 116, 295–305
[80]LeDoux, J., 2003. The emotional brain, fear, and the amygdala. Cellular and Molecular Neurobiology 23 (4–5), 727–738.
[81]Belova, M.A., Paton, J.J., Morrison, S.E., Salzman, C.D., 2007. Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron. 55 (6), 970–984.
[82]Baxter, M.G., Murray, E.A., 2002. The amygdala and reward. Nature Reviews Neuroscience 3, 563–572.
[83]Paton, J.J., Belova, M.A., Morrison, S.E., Salzman, C.D., 2006. The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature 439 (7078), 865–870.
[84]Lee, H.J., Groshek, F., Petrovich, G.D., Cantalini, J.P., Gallagher, M., Holland, P.C., 2005.Role of amygdalo-nigral circuitry in conditioning of a visual stimulus paired
with food. Journal of Neuroscience 25 (15), 3881–3888.

[85]Pare, Denis, and Sevil Duvarci. "Amygdala microcircuits mediating fear expression and extinction." Current opinion in neurobiology 22.4 (2012): 717-723.
[86]Joel, D., Weiner, I., 2000. The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience 96, 451.
[87]Gerfen, C.R., Herkenham, M., Thibault, J., 1987. The neostriatal mosaic. Ii. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. Journal of Neuroscience 7 (12), 3915–3934.
[88]Berendse, H.W., Groenewegen, H.J., Lohman, A.H., 1992. Compartmental distribution of ventral striatal neurons projecting to the mesencephalon in the rat. Journal of Neuroscience 12 (6), 2079–2103.
[89]Groenewegen, H.J., Berendse, H.W., Wolters, J.G., Lohman, A.H., 1990. The anatomical relationship of the prefrontal cortex with the striatopallidal system the thalamus and the amygdala: evidence for a parallel organization. Progress in Brain Research 85, 95–116 (discussion 116–118)
[90]Apicella, P., Scarnati, E., Ljungberg, T., Schultz, W., 1992. " Neuronal activity in monkey striatum related to the expectation of predictable environmental events. " Journal of Neurophysiology 68, 945–960
[91]Schultz, W., Apicella, P., Scarnati, E., Ljungberg, T., 1993b. Neuronal activity in monkey ventral striatum related to the expectation of reward. Journal of Neuroscience 12, 4595–4610
[92]Cromwell, H.C., Schultz, W., 2003. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. Journal of Neurophysiology 89, 2823–2838

[93]Ljungberg, T., Apicella, P., Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology 67, 145–163
[94]Mauk, M.D., Buonomano, D.V., 2004. The neural basis of temporal processing. Annual Review of Neuroscience 27 (1), 307–340
[95]Bentivoglio, M., Minciacchi, D., Morinari, M., Granato, A., Spreafico, R., Macchi, G., 1988. The intrinsic and extrinsic organization of the thalamic intralaminar nuclei. In: Bentivoglio, M., Spreafico, R. (Eds.), Cellular Thalamic Mechanisms. Excerpta Medica, Amsterdam, pp. 221–237.
[96]Lustig, C., Matell, M.S., Meck, W.H., 2005. Not just a coincidence: frontal-striatal
interactions in working memory and interval timing. Memory (Hove, England) 13, 441–448.
[97]Coull JT, Cheng R-K, Meck WH. " Neuroanatomical and Neurochemical Substrates of Timing. " Neuropsychopharmacology. 2011;36(1):3-25.
[98]Brown, J., Bullock, D., Grossberg, S., 1999. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. Journal of Neuroscience 19, 10502–10511.
[99]Sugrue LP, Corrado GS, Newsome WT. 2004"Matching behavior and the representation of value in the parietal cortex. " Science 304, 1782 –1787
[100]Platt ML, Glimcher PW. 1999" Neural correlates of decision variables in parietal cortex. " Nature400,233 –238.
[101]Shuler MG, Bear MF. 2006" Reward timing in the primary visual cortex. "Science 311, 1606 –1609.
[102]Vickery TJ, Chun MM, Lee D. 2011 "Ubiquity and specificity of reinforcement signals throughout the human brain. "Neuron72, 166 –177.

[103] Goltstein PM, Coffey EB, Roelfsema PR, Pennartz CM. 2013 "In vivo two-photon Ca2þ imaging reveals selective reward effects on stimulus-specific assemblies in mouse visual cortex. "J. Neurosci.33, 11 540 –11 555.
[104] Kilgard MP, Merzenich MM. 1998" Plasticity of temporal information processing in the primary auditory cortex. "Nat. Neurosci.1, 727 –731.
[105]O'Reilly, Randall C., Seth A. Herd, and Wolfgang M. Pauli. "Computational models of cognitive control." Current opinion in neurobiology 20.2 (2010): 257-261.
[106]Watanabe M. 1992 Frontal units of the monkey coding the associative significance of visual and auditory stimuli.Exp. Brain Res.89, 233 –247.
[107] Tremblay L, Schultz W. 1999 Relative reward preference in primate orbitofrontal cortex.Nature 398, 704 – 708.
[108] Padoa-Schioppa C, Assad JA. 2006 Neurons in the orbitofrontal cortex encode economic value.Nature 441, 223 –226.
[109]Kennerley SW, Dahmubed AF, Lara AH, Wallis JD.2009 Neurons in the frontal lobe encode the value of multiple decision variables.J. Cogn. Neurosci. 21, 1162– 1178
[110]Kouneiher F, Charron S, Koechlin E. 2009 Motivation and cognitive control in the human prefrontal cortex. Nat. Neurosci.12, 939 – 945.
[111]Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. 2005 The role of the dorsomedial striatum
in instrumental conditioning. Eur. J. Neurosci. 22, 513 –523.
[112]Alexander GE, Crutcher MD, DeLong MR. 1991 Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor,'prefrontal' and 'limbic' functions. Progress Brain Res.85, 119 –146.

[113]Tunstall MJ, Oorschot DE, Kean A, Wickens JR. 2002 " Inhibitory interactions between spiny projection neurons in the rat striatum. " J. Neurophysiol. 88, 1263 –1269.
[114]Taverna S, Van Dongen YC, Groenewegen HJ, Pennartz CM. 2004 " Direct physiological evidence for synaptic connectivity between medium-sized spiny neurons in rat nucleus accumbens in situ. " J. Neurophysiol. 91, 1111 – 1121.
[115]Redgrave P, Prescott T, Gurney KN. 1999 " The basal ganglia: a vertebrate solution to the selection problem? " Neuroscience89, 1009 –1023.
[116]Dickinson, A., Balleine, B.W., Watt, A., Gonzalez, F., and Boakes, 384–394. R.A. (1995). Motivational control after extended instrumental training. Anim. Learn. Behav. 23, 197–206
[117]Dayan, Peter, and Bernard W. Balleine. "Reward, motivation, and reinforcement learning." Neuron 36.2 (2002): 285-298.
[118]Nakamura, Kae. "The role of the dorsal raphé nucleus in reward-seeking behavior." Front Integr. Neurosci 7 (2013).
[119]Azmitia, E. C., and Gannon, P. J.(1986). The primate serotonergic system: a review of human and animal studies and a report on Macaca fascicularis. Adv. Neurol.43, 407–468.
[120]Daw, N. D., Kakade, S., and Dayan,P. (2002). Opponent interactions between serotonin and dopamine. Neural Netw. 15, 603–616.
[121]Zhou, Jingfeng, et al. "Prospective coding of dorsal raphe reward signals by the orbitofrontal cortex." Journal of Neuroscience 35.6 (2015): 2717-2730.

[122]Bechara A, Damasio H, Damasio AR (2000) "Emotion, decision making and the orbitofrontal cortex." Cereb Cortex 10:295–307.CrossRef Medline
[123]Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat Neurosci 1:155–159.CrossRef Medline
[124]Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. Nature 398:704–708.CrossRef Medline
[125]Wilson MA, Molliver ME (1991) The organization of serotonergic projections to cerebral cortex in primates: regional distribution of axon terminals. Neuroscience 44:537–553.CrossRef Medline
[126]Clarke HF, Dalley JW, Crofts HS, Robbins TW, Roberts AC (2004) Cognitive inflexibility after prefrontal serotonin depletion. Science 304:878–880.CrossRef Medline
[127]Molliver, M. E. (1987). Serotonergic neuronal systems: what their anatomic organization tells us about function. J. Clin. Psychopharmacol.7, S3–S23.
[128]Emergent Software,https://grey.colorado.edu/emergent/index.php/Main_Page
[129]Heathcote, Andrew, Scott Brown, and D. J. K. Mewhort. "The power law repealed: The case for an exponential law of practice." Psychonomic bulletin & review 7.2 (2000): 185-207.
[130]Ritter, Frank E., and Lael J. Schooler. "The learning curve." International encyclopedia of the social and behavioral sciences 13 (2001): 8602-8605.
[131]Collins, Anne GE, and Michael J. Frank. "How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic

analysis." European Journal of Neuroscience 35.7 (2012): 1024-1035.
[132]Hurlock, Elizabeth B., and Ella D. Newmark. "The memory span of preschool children." The Pedagogical Seminary and Journal of Genetic Psychology 39.2 (1931): 157-173.
[133]Miller, George A. "The magical number seven, plus or minus two: some limits on our capacity for processing information." Psychological review 63.2 (1956): 81.

[134]Cowan N.,The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci. 2001 Feb;24(1):87-114; discussion 114-85.

[135]Miyake, A. & Shah, P. (Eds.) (1999). Models of working memory. Mechanisms of active maintenance and executive control. Cambridge University Press
[136]Levin, E.S. (2011). Working Memory : Capacity, Developments and Improvement Techniques. New York: Nova Science Publishers, Inc.
[137]Cowan, Nelson (1995). Attention and memory: an integrated framework. Oxford [Oxfordshire]: Oxford University Press.
[138] Malenka RC, Nestler EJ, Hyman SE (2009). "Chapter 13: Higher Cognitive Function and Behavioral Control". In Sydor A, Brown RY. Molecular Neuropharmacology: A Foundation for Clinical Neuroscience (2nd ed.). New York: McGraw-Hill Medical. pp. 313–321.
[139] http://www.ni.com/it-it/shop/labview/download.html
[140]Jessell, Thomas M.; Kandel, Eric R.; Schwartz, James H. (2000). "27. Central visual pathways". Principles of neural science. New York: McGraw-Hill. pp. 533–540.

[141]LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).