A behavioral investigation of the algorithms underlying reinforcement learning in humans

Ana Catarina dos Santos Farinha

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Tiago Vaz Maia

Prof. Patrícia Margarida Piedade Figueiredo

Examination Committee

Chairperson: Prof. Cláudia Alexandra Martins Lobato da Silva

Supervisor: Prof. Tiago Vaz Maia

Member of the Committee: Prof. João Miguel Raposo Sanches

November 2014

“Believe you can and you're halfway there.”

— Theodore Roosevelt

Acknowledgments

First of all I would like to thank my supervisor Prof. Tiago Maia for all his support, guidance, and valuable insight throughout the last months. Also, I am truly grateful for the amazing group that I had the chance to work with (Ângelo, especially him, but also Ana, Inês, João, Lena, Rita and Vasco). Apart from the valuable discussions and lab meetings, the friendship and incentive that I found near them made this work possible.

I also need to thank all my friends for making me so spoiled. To the ones that are with me since ever or already for a long time (Daniela, Filipe Carvalho, Filipe Simões, João Borba, Sara, Sofia, Xana), to those that I recently met (like David, Duarte and Sofia Baeta), or even those far way (Esin, Jessica, Luigi and Karthiga). To Diogo Rechena, a special acknowledgment, for all his effort to help me during this study. And although many more fill my heart everyday one was exceptionally important through all the last five years: João Ramalhinho. Thank you all for everything.

Finally, but definitely not the least, I need to say how thankful I am for having the family that I have. Words cannot describe how lucky I feel. My parents and my brother are the best part of my life.

Abstract

Since the late 1990s, a large number of studies have been suggesting *reinforcement learning* (RL) as a computational framework within which decision-making can be explained. In fact, several lines of evidence link a key RL signal, the temporal difference reward prediction error, to the concentration of phasic dopamine. However, there is still not a clear understanding about how the neuronal circuits employed in learning processes work, and therefore which RL model better explains them. Moreover, RL has described several algorithms whose main difference is the way in which prediction errors are computed.

In this study a new probabilistic Go/NoGo task was designed and tested. The goal was to provide more insight regarding how prediction errors are computed in the human's brain: if using state values (like in the Actor-Critic model), or action values (as in the Q-learning model). The task addresses the question through both instrumental (Go/NoGo) and classical conditioning (subliminal images).

Behavioral data analyses from both approaches were in agreement, and in line with the Actor-Critic model. This conclusion suggests that humans follow this model, basing their decisions on prediction errors computed with state values.

Key Words

Actor-Critic, Q-learning, Prediction errors, Go/NoGo task, Dopamine.

Resumo

Desde o final da década de 90, que vários estudos têm sugerido a aprendizagem por reforços *(reinforcement learning)* como sendo uma estrutura normativa computacional na qual os processos de tomada de decisão podem ser explicados. De facto, várias evidências têm relacionado um sinal chave desta estrutura normativa, os erros de previsão utilizados nos algoritmos de diferenças temporais, com a concentração de dopamina fásica. No entanto, ainda não existe um conhecimento claro de como o circuito neuronal envolvido nos processos de aprendizagem funciona, e portanto qual o modelo computacional que melhor o explica. Além disso, a aprendizagem por reforços tem descrito vários algoritmos que diferem na maneira como os erros de previsão são calculados.

Neste estudo uma nova tarefa probabilística Go/NoGo foi desenhada e testada. O objectivo foi fornecer um maior conhecimento sobre o cálculo dos erros de previsão no cérebro humano: se determinados pelo valor do estado (como no modelo *Actor-Critic*), ou pelo valor da acção (como no modelo *Q-learning*). A tarefa aborda a questão através de condicionamento instrumental (Go/NoGo) e de condicionamento clássico (imagens subliminais).

A análise dos dados comportamentais de ambas as abordagens foi concordante e de acordo com o modelo *Actor-Critic*. Esta conclusão sugere que os humanos seguem este modelo, baseando as suas decisões em erros de previsão computados usando valores de estados.

Palavras Chave

Actor-Critic, Q-learning, Erros de previsão, Tarefa Go/NoG, Dopamina.

Contents

List of Figures

List of Tables

Abbreviations

AC	Actor-Critic
AI	Artificial Intelligence
ANOVA	Analysis Of Variance
BG	Basal Ganglia
BOLD	Blood Oxygen Level Dependent
CR	Conditioned response
CS	Conditioned stimulus
fMRI	functional Magnetic Resonance Imaging
GPe	Globus pallidus pars externa
GPi	Globus pallidus internal segment
MDP	Markov Decision Process
PE	Prediction Error
QL	Q-learning
RL	Reinforcement Learning
SNc	Substantia Nigra pars compacta
SNr	Substantia Nigra pars reticulada
STN	Subthalamic nucleus
SDT	Signal detection theory
TD	Temporal difference
UR	Unconditioned response
US	Unconditioned stimulus
VTA	Ventral Tegmental Area

List of Symbols

$\boldsymbol{t}$	discrete time step
$\boldsymbol{s_t}$	state at t
$\boldsymbol{a_t}$	action at t
$\boldsymbol{r_t}$	reinforcement at t, dependent on a_t and s_t
$\boldsymbol{R_t}$	return (cumulative discounted reinforcement) following t
$\boldsymbol{Q^{\pi}(s,a)}$	value of taking action a in state s under policy π
$\boldsymbol{\pi}$	policy, decision making rule
$\boldsymbol{\gamma}$	discount factor
$\boldsymbol{S}$	set of states
$\boldsymbol{\pi(s,a)}$	probability of taking action a in state s under policy π
$\boldsymbol{\widehat{Q}^{\pi}(s_t,a_t)}$	estimate of $Q^{\pi}(s,a)$
$\boldsymbol{A(s_t)}$	set of possible actions in s_t
$\boldsymbol{p(s_t,a_t)}$	preference of taking action a in state s under policy π
$\boldsymbol{\widehat{V}^{\pi}(s_t)}$	estimate of $\boldsymbol{V^{\pi}(s)}$
$\boldsymbol{V^{\pi}(s)}$	value of state $\boldsymbol{s_{t+1}}$ under policy $\boldsymbol{\pi}$
$\boldsymbol{\alpha}$	learning rate
$\boldsymbol{\beta}$	inverse of temperature
$\boldsymbol{d'}$	sensitivity index
$\boldsymbol{P^{a}_{ss'}}$	probability of transition from state $\boldsymbol{s}$ to state $\boldsymbol{s'}$ under action $\boldsymbol{a}$
$\boldsymbol{R^{a}_{ss'}}$	expected immediate reward on transition from $\boldsymbol{s}$ to $\boldsymbol{s'}$ under action $\boldsymbol{a}$

Glossary

Action-value – returns the value, i.e. the expected return for using action a in a certain state s. Return means the overall reward.

Associative learning – Process by which an association between two stimuli or a behavior and a stimulus is learned. The two forms of associative learning are classical and operant conditioning.

Behavioral data - Observational reports about the behavior of organisms and the conditions under which the behavior occurs or changes.

Blocking - A phenomenon in which an organism does not learn a new stimulus that signals an unconditioned stimulus, because the new stimulus is presented simultaneously with a stimulus that is already effective as a signal.

BOLD – functionalMRI technique that uses the differences in magnetic susceptibility between oxyhemoglobin and deoxyhemoglobin to image areas of activated cerebral cortex.

Classical conditioning - A type of learning in which a behavior (conditioned response) comes to be elicited by a stimulus (conditioned stimulus) that has acquired its power through an association with a biologically significant stimulus (unconditioned stimulus).

Conditioned reinforcers - In classical conditioning, formerly neutral stimuli that have become reinforcers.

Conditioned response (CR) - In classical conditioning, a response elicited by some previously neutral stimulus that occurs as a result of pairing the neutral stimulus with an unconditioned stimulus.

Conditioned stimulus (CS) - In classical conditioning, a previously neutral stimulus that comes to elicit a conditioned response.

Conditioning - The ways in which events, stimuli, and behavior become associated with one another.

Confounding variable - A stimulus other than the variable an experimenter explicitly introduces into a research setting that affects a participant's behavior.

Consciousness - A state of awareness of internal events and of the external environment.

Decision making - The process of choosing between alternatives; selecting or rejecting available options.

Discount factor - A scalar value between 0 and 1 which determines the present value of future rewards. If the discount factor is 0, the agent is concerned with maximizing immediate rewards. As the discount factor approaches 1, the agent takes more future rewards into account.

Fixation.- A state in which a person remains attached to objects or activities more appropriate for an earlier stage of psychosexual development.

functional MRI (fMRI) - A brain imaging technique that combines benefits of both MRI and PET scans by detecting magnetic changes in the flow of blood to cells in the brain.

Habits - S-R associations and function almost like reflexes.

Instrumental conditioning - Learning in which the probability of a response is changed by a change in its consequences.

Ion channels - The portions of neurons' cell membranes that selectively permit certain ions to flow in and out.

Law of effect - A basic law of learning that states that the power of a stimulus to evoke a response is strengthened when the response is followed by a reward and weakened when it is not followed by a reward.

Learning - A process based on experience that results in a relatively permanent change in behavior or behavioral potential.

Neuromodulator - Any substance that modifies or modulates the activities of the postsynaptic neuron.

Neuron - A cell in the nervous system specialized to receive, process, and/or transmit information to other cells.

Neuroscience - The scientific study of the brain and of the links between brain activity and behavior.

Neurotransmitters - Chemical messengers released from neurons that cross the synapse from one neuron to another, stimulating the postsynaptic neuron.

Pain - The body's response to noxious stimuli that are intense enough to cause, or threaten to cause, tissue damage.

Perception - The processes that organize information in the sensory image and interpret it as having been produced by properties of objects or events in the external, three-dimensional world.

Positive reinforcement - A behavior is followed by the presentation of an appetitive stimulus, increasing the probability of that behavior.

Policy – a mapping from stimuli to responses on the basis of the previous reward history.

Primacy effect - Improved memory for items at the start of a list.

Punisher - Any stimulus that, when made contingent upon a response, decreases the probability of that response.

Reinforcement Learning – Branch of Artificial Intelligence that focuses on learning from interactive experiences. Also used to describe the collection of processes whereby humans and animals learn though rewards.

Reasoning - The process of thinking in which conclusions are drawn from a set of facts; thinking directed toward a given goal or objective.

Recency effect - Improved memory for items at the end of a list.

Recognition - A method of retrieval in which an individual is required to identify stimuli as having been experienced before.

Reinforcer - Any stimulus that, when made contingent upon a response, increases the probability of that response.

Response bias - The systematic tendency as a result of nonsensory factors for an observer to favor responding in a particular way.

State-value - returns the value, i.e. the expected return for selecting a certain state s. Return means the overall reward.

Safety Signal - Safety signals are learned cues that predict the nonoccurrence of an aversive event. As such, safety signals are potent inhibitors of fear and stress responses.

Serial position effect - A characteristic of memory retrieval in which the recall of beginning and end items on a list is often better than recall of items appearing in the middle.

Signal detection theory (SDT) - A systematic approach to the problem of response bias that allows an experimenter to identify and separate the roles of sensory stimuli and the individual's criterion level in producing the final response.

Significant difference - A difference between experimental groups or conditions that would have occurred by chance less than an accepted criterion; in psychology, the criterion most often used is a probability of less than 5 times out of 100, or $p < .05$.

Stimulus-response (S-R) learning – The learning of an association between a stimulus and a response, with the result that the stimulus comes to elicit the response.

Stimulus-stimulus (S-S) learning – The learning of an association between two stimuli, with the result that exposure to one of the stimuli comes to activate a representation, or "mental image", of the other stimulus.

Subliminal perception – Perception below the threshold or limen of consciousness.

TD (temporal difference) algorithms - A class of learning methods, based on the idea of comparing temporally successive predictions.

Three-term contingency - The means by which organisms learn that, in the presence of some stimuli but not others, their behavior is likely to have a particular effect on the environment.

Unconditioned response (UR) - In classical conditioning, the response elicited by an unconditioned stimulus without prior training or learning.

Unconditioned stimulus (US) - In classical conditioning, the stimulus that elicits an unconditioned response.

Nonconscious - Information not typically available to consciousness or memory.

1. Introduction

Contents

1.1 Motivation

One of the most important aspects of our lives is the way in which we as humans, and more generally as animals, adapt to the surrounding environment. This adaptation is supported by learning appropriately how to behave in each situation: if an action leads to a reward (such as food), then that action should be reinforced so that in similar future situations the animal can choose it more efficiently; on the other hand, if an action leads to a punishment (for instance, a shock), then it should be avoided. The idea of decision making as a response to rewards and punishments has been supported with several animal behavior studies, leading to the development of classical and instrumental conditioning theories [1].

So, if we consider that punishments act as negative rewards signals, it is not surprising that artificial intelligence development soon tried to create algorithms that, making use of this reward signal, led to *self-learning* machines [2, 3]. In fact, great progresses have been made in this area, called *reinforcement learning* (RL) throughout the years. Moreover, multiple lines of evidence have been linking the RL framework to the function of dopaminergic neurons in the mammalian midbrain (mostly by extracellular recordings during behavior tasks) [4-6] and, more recently, to data from human decision-making imaging experiments in fMRI (**f**unctional **M**agnetic **R**esonance **I**maging) cameras [7-9]. A key link that arose is that dopamine appears to be the correlated with a key RL signal, the *temporal difference reward prediction error* (TD PE) [10-13].

This combination of computer science and neuroscience has the potential to make important contributions to both areas:

- A better understanding of the neural mechanisms of human decision-making can be a source of information of how RL algorithms should be designed in order to improve artificial systems capable of dealing with real-world environment;
- A deeper knowledge about the neural processes in the normal and pathological brain is the first step to change the current symptom-based classification system of mental disorders into a system based on pathophysiology [12]. For instance, one important neurocomputational model was inspired in the corticostratio-thalamocortical loops, the basal ganglia Go/NoGo (BG-GNG) model [14]. Modeling how these loops work in the normal brain and in disorders caused by to disturbances in those loops (such as Parkinson's disease [14-17], Tourette's syndrome [17, 18], and many others) can help reversing or treating these conditions in a better way.

However, various types of RL algorithms have been developed - for instance prediction errors can be calculated by different forms: using the value of actions (Q-learning model [2, 3, 19]) or the value of states (Actor-Critic model [2, 19, 20]). Since there is neuroscientific data corroborating both

models, there is an ongoing discussion about which model better mimics our brain's functions. Therefore, it is of utmost importance to test more complex scenarios of human decision-making behavior, as they might provide a better insight regarding how and which RL method actually should be implemented.

1.2 Objective

The aim of this study is to improve the knowledge about how prediction errors are computed in the human brain, namely if they are determined by state values (as in the Actor-Critic model) or by action values (as in the Q-learning framework). For that, we intent to design a new Go/NoGo task (particularly focused on addressing this controversy) test it in a fairly number of subjects and analyze the behavioral data to draw new insights.

1.3 Thesis Outline

The present work is divided into four main parts:

i. It starts by giving the reader the necessary background to understand the problem under debate. Firstly, concepts, findings and theories regarding psychological studies are described; then some basic neuroscience concepts (how and where learning occurs) are also covered , followed by notions in computer science theory; and finally evidences that largely contributed to the synergy between the just referred areas are presented;

ii. The detailed description of the design of the new Go/NoGo task developed, and the reasoning behind all the choices made;

iii. In this section the results of the behavioral analysis are shown. Each of the constituent phases of the designed task will be consider in turn;

iv. Finally, the last section summarizes the study key remarks and conclusions, and presents ideas for future improvements.

2. Background

Contents

In this chapter some background information regarding psychological behaviour studies, basic concepts about learning neuroscience and computer science will be exposed. Also, evidences supporting the synergy between the aforementioned areas are shown. This will give the reader the necessary foundations to understand the question addressed in this study.

2.1 Psychological studies

Psychological behaviorism is a major theory in psychology which holds that behaviors are learned through positive and negative reinforcements. Studies in this filed can be divided into classical and instrumental conditioning and the main different between these two forms is the interactive aspect of the latter one (the subject needs to perform an action to have a reward, while in the former form he remains passive).

In this section both types of conditioning will be explained.

2.1.1 Classical Conditioning

The genesis of the **classical conditioning** paradigm dates back to the beginning of the twentieth century with Pavlov's work. While studying processes of digestion – an area that awarded him the Nobel Prize in Physiology, in 1904 – he noticed that dogs secreted stomach juices on response to the sight of food, or even just upon seeing the person who usually fed them. He named those stomach secretions elicited by food-related stimuli as *psychic secretions* and recognized their potential, suggesting that they could be studied with special attention to reveal the mechanisms of association learning – the process by which an association between two stimuli or a behavior and a stimulus is learned.

The basic procedure of the experiments conducted afterwards is well-known. It involves two stimuli: 1) a conditioned stimulus (**CS**), such as light or a tone that does not elicit salivation at the outset of the experiment, and 2) an unconditioned stimulus (**US**) like delicious food or sour-taste that does elicit salivation without any training, and thus generate an unconditioned response (**UR**). After pairing several times the CS with the presentation of food, it also started eliciting salivation, generating a conditioned response (**CR**) [1].

One of the critical questions that aroused was to explain how a CS produces a response. Two different approaches were suggested:

- In one approach, conditioned behavior is viewed as a response elicited directly by the CS through establishing a new *stimulus-response* (S-R) connection between the CS and the CR;

- An alternative view is that subjects do not learn a new S-R connection but instead a new *stimulus-stimulus* (S-S) connection since, according to this theory, the CS does not elicit a CR directly, but instead activates a representation or memory of the US.

Several studies have supported the latter theory, whereby the **US-devaluation** technique is one of the most acceptable arguments in this debate [21].

To better understand the devaluation technique, an **US-devaluation experiment** is presented [21] (fig. 2.1.):

- In Phase 1 there are two groups of mildly food-deprived rats, conditioned by repeatedly pairing a tone with pellets of food – consequently, an association between the CS (tone) and the US (food) is formed;
- In Phase 2, in the experimental group, and contrarily to what happens in the control group, the US representation is devaluated by giving sufficient free food to the rats so that they can completely satisfy their hunger (decreasing food's value);
- In the Test phase, both groups receive a series of test trials with the CS (tone);

The results show a decrease in conditioned responding for the experimental group. This decrease supports the idea that the presentation of the CS activates the US representation, and that the CR is dependent on the current status of that US representation (otherwise CS would have elicited the same CR whenever it occurred regardless of the food value).

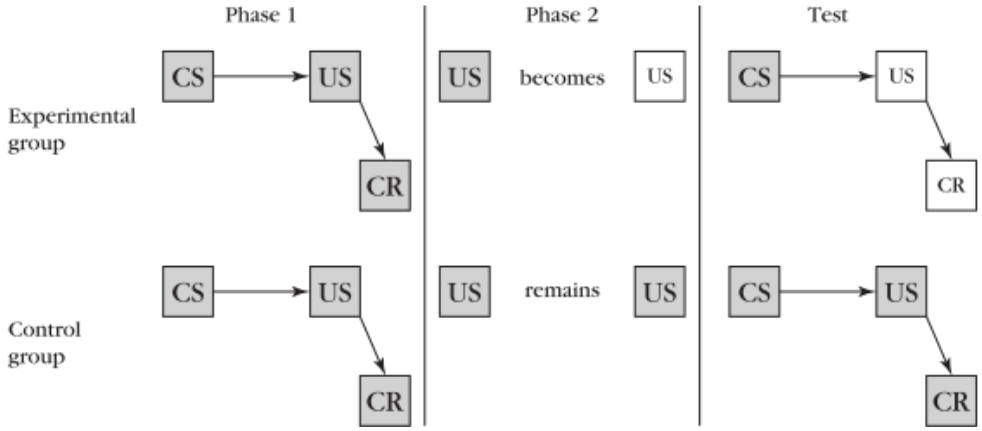


Figure 2.1: Basic strategy and rationale involved in US-devaluation experiments. In Phase 1 the experimental and control groups receive conventional conditioning to establish an association between the CS and the US and to lead the participants to form a representation of the US. In the Phase 2 the US representation is devaluated for the experimental group but remains unchanged for the control group. If the CR is elicited by way of the US representation, devaluation of the US representation should reduce responding to the CS (from [1]).

Another important aspect revealed by these experiments is the idea that learning only happens for surprising or unexpected rewards. For instance, if during an experiment a CS is paired with a reward until the association is learned, but then a second CS is introduced (being now both CSs paired with the reward), the second CS alone will not elicit conditioned responding. This happens because when the newly introduced CS was paired, the US was not surprising anymore (since the first CS already predicted it), and so the S-S connection described above will not be formed for that new CS. This concept, referred to as the **blocking effect** (shown on multiple studies), was formally expressed by Robert Rescorla and Allan Wagner in 1972 [22], and their model (**RW model**) has been a reference ever since. By denoting the US value as λ (what is received on a given trial) and V the **associative value** of the CS, the model can be expressed as:

$$\Delta V = k(\lambda - V) \tag{2.1}$$

Being $(\lambda - V)$ the difference between what occurs (λ) and what is expected (V), and k a constant related to the salience of the CS and US that controls the speed of learning. It can be easily understood that with the progression of the learning process the surprise term will be smaller, converging to zero when the learning is fully accomplished [1].

2.1.2 Instrumental Conditioning

In the class of conditioning behavior previously described the subject does not need to perform any particular response in order to obtain the UR or the CR. For example, the dog in the famous Pavlov's experiment did not do any particular action in order to receive the food. Thus, it reflects how organisms adjust to events in their environment that they cannot directly control. On the other hand, in **instrumental conditioning,** responding/performing an action is necessary to produce a desired environmental outcome. For instance, a student knows that by studying hard can achieve a better grade in an exam. In this example, the action studying is **instrumental behavior,** which is done in order to produce certain expected outcomes.

One of the pioneers in this area was the American psychologist Edward L. Thorndike with his famous puzzle boxes studies. His training procedures consisted of putting a hungry animal inside a puzzle box with some food outside, but visible to him. This setup forced the animal to learn how to get out of the box and reach the food. It was observed that the animal became faster and faster in escaping the puzzle box. Thorndike's interpretation was that the animal was not actually gaining insight about the puzzle itself (and so learning the release mechanism), but by using a trial-and-error approach the animal was learning an S-R association. In 1911, Thorndike formulated his famous ***Law of Effect*** [23]*:*

> *"Of several responses made to the same situation, those which are accompanied by or closely followed by satisfaction to the animal will, other things being equal, be more firmly*

connected with the situation, so that, when it recurs, they will be more likely to recur…" (p.244)

This law states that when a response in presence of a stimulus is followed by a *satisfying* outcome, the S-R connection should be fortified. And, oppositely, when it is followed by an undesirable outcome, then the S-R connection should be weakened. Nevertheless, it should be noted that the outcome of the response is not taken into account for the association process [3, 24]. This led to another instrumental conditioning theory that establishes a link between response and outcome (R-O) [24].

An empirical argument in favor of the R-O association comes from **reinforcer devaluation** experiments (e.g. [25, 26]). Similar to the US-devaluation experiments described before, the experiments start with a conditioning training, but on this case by forcing an animal to press a lever in order to get some reward (food). Afterwards, in the experimental group the food is devaluated (as before, by giving enough food to make the animal satisfied). Finally, during the test phase, both the animals from the experimental group and from the control group are placed again in the experimental set-up. Results show that animals from the former group press the lever less than the ones from the control group. This decrease in responding provides evidence that the animals consider the knowledge about the outcome, and do not act only due to a reflex to the stimulus.

Taking into account these results, the S-R learning is usually related to **habits** and is a devaluation-insensitive behavior, whereas the R-O learning is also called **goal-directed** behavior and is devaluation-sensitive. To understand the differences more clearly another example is presented, shown in figure 2.2 [27].

Imagine a scenario where one has just left work, on a Friday evening, and needs to decide which route to take on the way back home. The problem can be thought as having several *states* (here locations), *actions* (such as going straight or turning left), *probabilities of transitioning* from one state to another when a certain action is performed (that do not need to be deterministic, since unpredictable events can happened), and positive and negative *outcomes* when an action is taken (again, probabilistic).

So, the problem of deciding the route can be thought in two different ways [27]:

- In the first one, exemplified by the left side of the figure 2.2, a mental map has been learned and a **model-based** computation can be performed. Since the peculiarities of the task are already known (such as its transitions probabilities and rewards probabilities - concepts further explained in section 2.3), planning for the route is a matter of choosing the one with the highest value. The model-based approach is a goal-direct behavior and changes in the environment cause an immediately shift in subject's behavior.

- On the other hand, if we perform a habit, it leads us to the **model-free** action selection area (right side of the figure). Here we do not make use of any constructed mental model but instead use values that reflect stored experiences that denote the overall future worth of an action. We could, for example, know by experience that taking the freeway on a Friday evening is not a good option, and thus the value of this action is lower when compared to the option of going straight in that intersection. Because values are only good estimates of future consequences with experience when a change in the environment occurs (e.g. unexpected traffic in a particular road) the subject is not able to adapt rapidly to it – he needs to experience it in order to update the overall future worth of actions in that new situation.

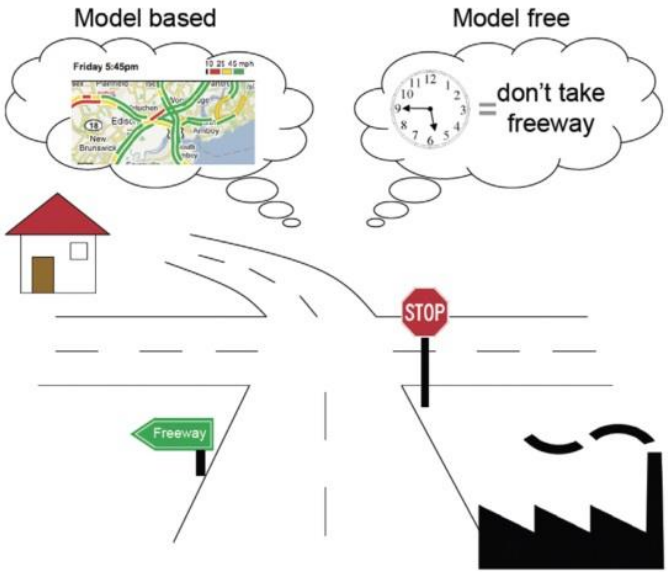


Figure 2.2: Two ways to choose which route to take when traveling home from work. Model based (habit) versus Model free (goal-directed) behavior (from [27]).

Several studies have suggested that our brain implements both strategies in parallel, being one or the other dominating depending on the circumstance (e.g. [28]). Moreover, different neural substrates might underlie each one of these approaches (e.g. [29-31]).

2.2 Neuroscience

Before providing more details about how learning is processed in the human brain, it is important to clarify some basic concepts about the neurobiology and the anatomy of the brain (more specifically of the basal ganglia, deemed the most important structures involved in learning).

Communication in the brain is made through *action potentials*. Action potentials are rapid changes in the membrane potential (difference in electric potential between the interior and the exterior of a biological membrane) that spread rapidly along the nerve fiber membrane (typically with 100 mV of amplitude, 1 ms of duration, and with a velocity of conduction along the axon that ranges from about 1 to 100 m/s depending on the axon diameter). Each of them starts with a sudden change from the normal negative membrane potential, the **resting stage**, to a positive potential (*depolarization*) and then ends with an almost equally rapid change back to the negative potential (*repolarization*). To conduct a nerve signal, the action potential moves along the nerve fiber until it comes to its end [32, 33] .

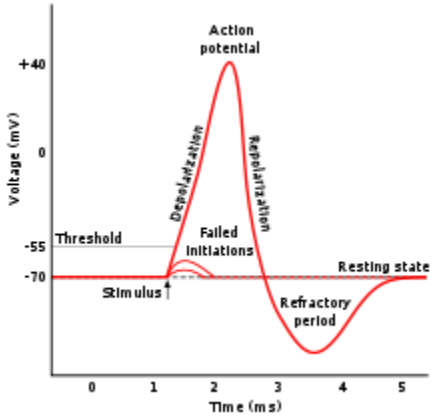


Figure 2.3: Successive steps during an action potential.

As long as the nerve's fiber membrane remains undisturbed, no action potential occurs. However, when an event pushes the membrane potential from the resting (negative) value towards zero, this rising voltage itself leads to the opening of multiple sodium voltage-gated ion channels (Na^+). Voltage-gated ion channels are channels that allow ions to move into and out of the cell controlling the flow of ions across the cell membrane by opening and closing in response to voltage changes and to both internal and external signals. This allows for a rapid inflow of sodium ions, which further reinforces the increase of the membrane potential value, leading to the opening of additional voltage-gated sodium channels. This configures a positive-feedback cycle that, once the feedback is strong enough, continues until all the voltage-gated sodium channels have become active (open). Once the membrane potential reaches its peak, the sodium channels begin to close and the potassium channels begin to open (K^+), terminating the action potential. The combined effect of a decrease in the entrance of sodium ions and the simultaneous increase in the exit of potassium ions accelerates the repolarization process, leading to the full recovery of the resting membrane potential [32, 33].

Recording the brain's electrical activity during behavior tasks has greatly contributed to our knowledge of the learning process. Nevertheless, it should be noted that the action potentials are

not recorded directly on the cell, but as voltages changes outside (but close) to the body of a nerve cell – as extracellular recorded potentials (spikes). This information is usually presented as a raster plot in which the occurrence of each spike during a trial is represented by a dot along the time axis. Data from multiple trials are aligned horizontally and separated vertically, as can be seen in figure 2.4. [34]

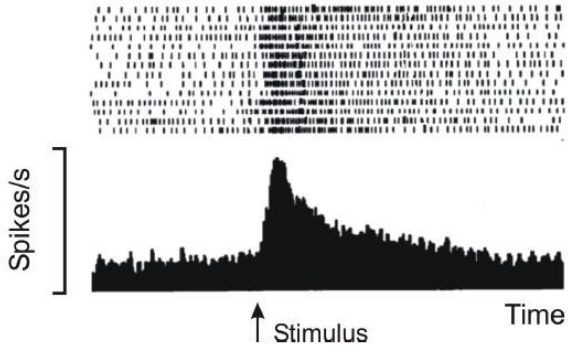


Figure 2.4: In the top: a raster plot showing recordings from multiple trials. Trials are aligned on the time of the stimulus. At the bottom: histogram showing the average activity across all trials (from [34]).

The communication between neurons is accomplished through the release of chemical transmitters (**neurotransmitters**) from the axon terminals of one neuron (the presynaptic neuron) to sites localized on the dendrites and/or cell body of another neuron (the postsynaptic neuron) through synapses. Neurotransmitters will bind to receptor molecules associated with ion channels and change their ionic conductance (within less than a millionth of a second since the release). Depending on the transmitter's type and on the type of the postsynaptic receptor, the postsynaptic stimulation can be excitatory (increasing its activity) or inhibitory (decreasing its activity).

Besides neurotransmitters, **neuromodulators** (such as dopamine, noradrenaline, serotonin and acetylcholine) are also released to the synaptic cleft between two different neurons. However, they operate in a different way, as neuromodulators bind to receptors coupled to membrane proteins (G-proteins). These G-proteins are connected to other membrane molecules that, when being activated, increase the level of molecules called second messengers inside the postsynaptic neurons and/or axonal terminals. Secondary messengers are associated with several functions and can, for example, increase the strength of a synaptic connection (a requirement for learning) [33].

2.2.1 How does learning take place in our brain? Basic cellular neurobiology concepts

The learning process (as well as other processes like those related to memory), requires long-term modifications in the brain's neuronal networks. The capacity of neuronal networks to change

with experience is called *neuronal plasticity*. The idea that learning is mostly related to activity-dependent synaptic plasticity was formulated in the *Hebb rule*, in 1949 [35]:

> *" When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."*

This suggests that such synaptic modifications could produce new neuronal connections that reflect the relationships learned during training. Consequently, if some neurons fire together in consequence of an association between a stimulus and a response, it is a desirable trait that these neurons are able to develop stronger interconnections so that, in future similar situations, when a stimulus activates some of the associated neurons, a synaptic drive would activate the remaining neurons related to the appropriate response.

Latter, studies from Jeff Wickens and colleagues found that the plasticity of cortico-striatal synapses is also weighted by the dopamine input: when presynaptic and postsynaptic activation is associated with increased dopamine input (long-term potentiation, **LTP**), the synaptic connection is strengthened; whereas, if the presynaptic and postsynaptic activation is not associated with dopamine input (long-term depression, **LTD**), the connection is depressed [12, 36-38].

2.2.2 Where do the modifications happen? Anatomy of the Basal Ganglia

From a global perspective, the brain is formed by several major structures that include the cerebral hemispheres, the brain stem and the cerebellum. Accumulated evidence supports the idea that distinct nuclei, the basal ganglia, play a particularly important role in the learning process (mainly those based on reinforcers). This is where dopaminergic modulation occurs, allowing for the emergence of a wise and intelligent behavior face to the environment and based on previous events [15, 39, 40].

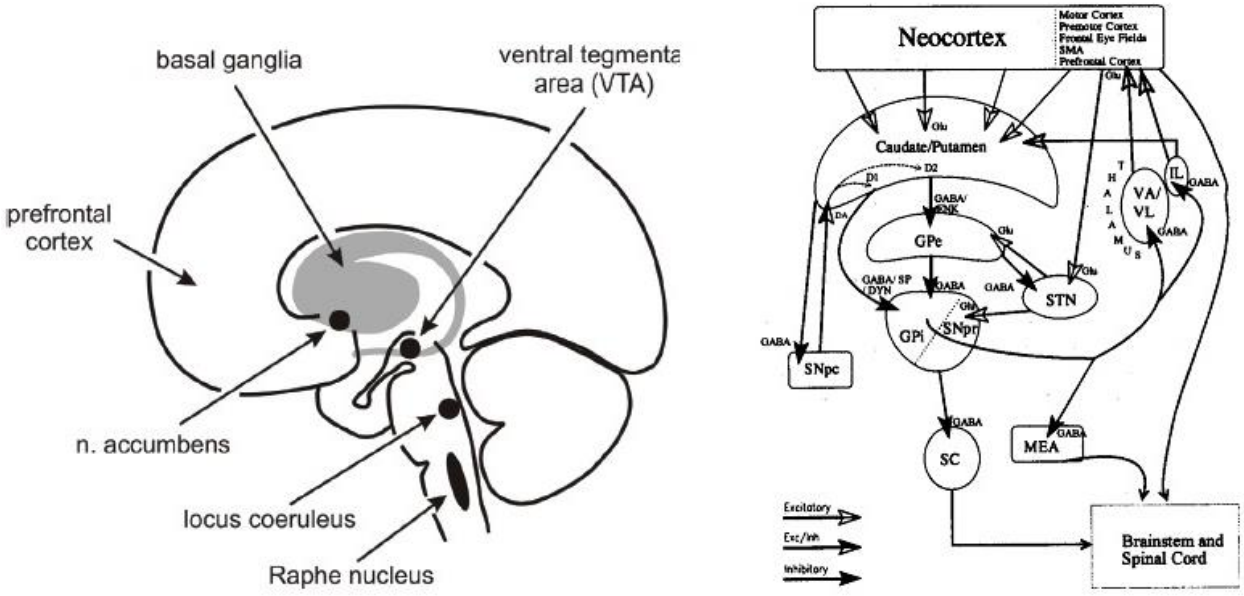


Figure 2.5: In the left: drawing of a sagittal section through the brain showing some stem nuclei (in black) and the basal ganglia (in grey). In the right: Diagram of the basal ganglia showing some of the most important dopaminergic pathways (from [34] and [39]).

The basal ganglia are a group of interconnected subcortical nuclei distributed across several main brain areas: the telecenphalon, the diencephalon, and the midbrain; they can be divided into: i) afferent structures, ii) output structures and iii) intrinsic nuclei [39].

i) The primary **afferent structure** is the striatum (also known as neostriatum), and it is formed by the caudate, putamen, and nucleus accumbens. This structure receives input from virtually all areas of the cerebral cortex, but sends output only to the other components of the basal ganglia. One aspect that is worth highlighting is the fact that most cortical inputs terminate in dorsal regions of the striatum, while inputs from the brain stem nuclei terminate in both dorsal and ventral areas of the striatum – this difference can be due to dissociable roles of ventral and dorsal striatum in instrumental conditioning [41]. The other input structure of the BG is the subthalamic nucleus (STN) that receives input mainly from motor areas of the frontal lobe;

ii) The **output structures** are composed by the globus pallidus internal segment (GPi) and the substantia nigra pars reticulata (SNr), which are mainly involved in movement. These inhibitory structures project to motor areas in the brainstem and thalamus in order to generate and control purposive movements. As can be seen in figure 2.5, both structures receive excitatory input from the STN and inhibitory input from the striatum;

iii) The **intrinsic nuclei** is comprised of the globus pallidus pars externa (GPe) and the substantia nigra pars compacta (SNpc). The former receives inhibitory input from the striatum and excitatory input from the STN, and projects in an inhibitory way to the STN and to the output structures. The latter one is the locus of dopamine-containing neurons and connects primarily to the striatum.

2.2.3 How do the corticostriatal-thalamocortical loops in BG work?

It is broadly agreed that basal ganglia is primarily implicated in selecting the best action to execute it at a given time. Its output facilitates the execution of a single motor command and inhibits competing motor mechanisms so that movement can proceed without interference. This is accomplished through the balance of two main output pathways: i) the direct (Striatonigral or Go pathway) and the ii) indirect (Striatopallidal or NoGo pathway) [42].

i) Via the **direct pathway**, the striatum inhibits directly the output structures (the SNr and the GPi), leading to the disinhibition of the thalamus. This disinhibition allows for the occurrence of the appropriate movement;

ii) Via the **indirect pathway**, the striatum influences the output structures through the inhibition of the GPe. Consequently, the output structures will no longer be disinhibit, which results in inhibition of the thalamus and thus constrainment of movements. This pathway can also involve the subthalamic nucleus (STN) since the GPe can inhibit it, resulting in an excitatory influence on the GPi/SNr through glutamatergic connections.

There is still a third cortico-basal ganglia pathway, the **hyperdirect pathway**. In this pathway, the mediofrontal cortex excites the STN, which then provides excitatory drive to the output nuclei of the basal ganglia. This pathway has been shown to have particular relevance under conditions associated with response conflict [14] and preventing premature behavior [12, 43].

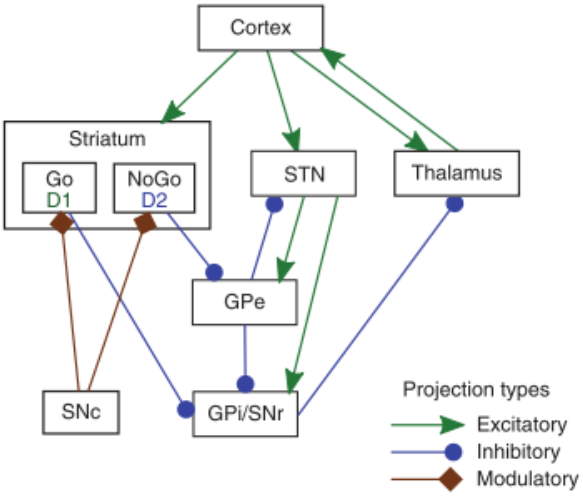


Figure 2.6: Diagram of the Go/direct pathway and NoGo/indirect pathway (adapted from [12]).

2.2.4 Dopamine receptor regulation of the striatal-projection pathways

Even though the neurons in the direct and the indirect pathways share some common morphological and neurochemical characteristics, they can be distinguished not only by their projections, but also by their expression of different neuropeptides and dopamine receptor subtypes.

There are five types of G protein-coupled dopamine receptors and they can be divided into two main groups: D1 and D2, based on how they respond to agonists [39, 40]. Evidences, primarily from *in situ* hybridization histochemistry, verified that most striatopallidal neurons express the peptide enkephalin and the D2 dopamine receptor, whereas most striatonigral neurons express both substance P and dynorphin and the D1 dopamine receptor [44, 45]. This difference is the main responsible factor for the opposing effects that dopamine exerts on these striatal output pathways. Therefore, understanding it gives us insight into how cortical input to the striatum is processed to affect its output neurons.

At rest, the striatal spiny neurons are physiologically quiescent (silent), whereas the neurons of the output structures are tonically active [40]. When the direct pathway is active, GO neurons phasically inhibit the tonic activity of nigral neurons, leading to the disinhibition of the thalamus. On the other hand, when the indirect pathway is active, there will be a disinhibition of the subthalamic nucleus, increasing the tonic firing of nigral neurons and so, inhibiting the thalamus' activity [12, 14].

In cases of a dopamine-depleted striatum (either due to lesions of the nigrostriatal dopamine pathway or in the dopamine neurons of the SNc, for example) the absence of the stimulatory effect that dopamine exerts through the D1 receptor leads to a decrease of the expression of substance P and dynorphin, in the direct pathway [40]. Oppositely, in the indirect pathway, the absence of

dopamine-mediated inhibition through the D2 receptor causes an increase of markers like enkephalin [40].

2.3 Computer science

The observations given by psychological investigations of conditioned behavior were a booster for artificial intelligence (AI), which tries to mimic them in machine learning computer algorithms. In this chapter, basic concepts about reinforcement learning algorithms are described, especially the ones where temporal-difference learning is implied (such as SARSA, Q-learning (QL) and the actor-critic (AC)), are presented.

2.3.1 Reinforcement learning: Temporal Difference learning (TD)

Reinforcement learning is a branch of AI that focus on learning from interactive experience. During the learning process, the agent (the decision-maker), interacts with the environment through a sequence of discrete time steps ($t = 0,1,2$ ...). At each moment he is presented with a situation, a state s_t (from a set S of all possible states $s_t \in S$) and, according to a policy π_t (which is basically a mapping from stimulus to responses), he makes an action a_t from a set of actions $A(s_t)$ - avaiable in that state ($a_t \in A(s_t)$).

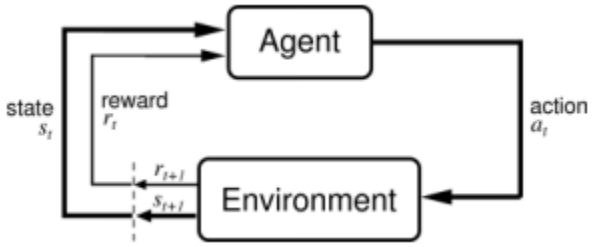


Figure 2.7: The agent-environment interaction in RL (from [34]).

The policy $\pi_t(s, a)$ is what the agent needs to improve to achieve his goal: maximize rewards and diminish punishments given the surronding environment. However, in most daily situations his actions may affect not only the immediate reward but also future rewards – this is the key difference between TD learning and Rescorla and Wagner's framework where associative strenghts only consider the immediately forthcoming reward, being a timeless entity. We can consider the expected reward (or return) at a given time as a sum of rewards:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \quad (2.2)$$

Where γ is a parameter called discount rate ($0 \leq \gamma \leq 1$). This parameter makes sure that the infinite sum has a finite value, as long as the reward sequence $\{r_t\}$ is bounded, and determines the present value of future rewards, meaning that a distant reward will not be valued as much as the

ones received immediately. This also agrees with the fact that humans and animals prefer earlier rewards to later ones.

Another important assumption in reinforcement learning problems is that in a given state the future is entirely independent of the history before that state. This assumption ensures that the knowledge of the current state is sufficient to predicti anything that can be known about the future. This property is designated as the *Markov property* and a state that follows it is denominated as *Markov*. One example, described by Richard Sutton and Andrew Barto [2] is that a checkers position is a Markov position because even though much of the information about the sequence of events before that state is lost, it summarizes all the important information need for the player to make his decision about the future.

A reinforcement learning task that satisfies the Markov property is a *Markov decision process* or *MDP*. Moreover, if the state and action spaces are finite, then, it is a *finite Markov decision process* (*finite MDP*).

An individual *finite MDP* is defined by its state and action sets and by the one-step dynamics of the environment. The probability that in a given state s performing an action a will lead to the next state *s'* is translated into quantities called *transition probabilities*:

$$P_{ss'}^{a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{2.3}$$

In the same way, we define the *expected value of the next reward* as:

$$R_{ss'}^{a} = \mathrm{E}\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \tag{2.4}$$

These two quantities configure the most important aspects of the dynamics of a finite MDP.

As mentioned previously, when the agent chooses an action to perform, he needs to take into account not only the immediately reward but also *how good* it is for him to be in a given future state s_{t+1}, which reflects the expected value of the discounted sum of future rewards that the agent will receive next. This implies by the agent the computation of a *state-value* function for a given state s under a policy π denoted as $V^{\pi}(s)$. For MDPs, the state-value function is denoted as:

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\textstyle\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\} \tag{2.5}$$

Where $E_{\pi}\{\}$ represents the expected value given that the agent follows policy π at time t. Similarly, we can define an *action-value function for policy* π, $Q^{\pi}(s,a)$ as the value of taking action a in state s under the policy π:

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\{\textstyle\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\} \tag{2.6}$$

The problem that stands out now is defining how the agent actually performs the computations of the state-value function and of the action-value function. If the agent knows the dynamics of the reinforcement learning problem, i.e. the transition probabilities and the expected value of the next reward, solving the MDP is straightforward:

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$
$$= E_{\pi}\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s\}$$
$$= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_{t+1} = s'\right\}$$
$$= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s')] \quad (2.7)$$

And in a similar way:

$$Q^{\pi}(s,a) = r_{t+1}(a_1) + \gamma r_{t+2} + \cdots = E_{\pi}\{R_t | s_t = s, a_t = a\}$$
$$= r_{t+1}(a_1) + \gamma V(s_{t+1}) = \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s') | s_{t+1} = s'] \quad (2.8)$$

The key difference between action-value functions and state-value functions is that the former do not depend on all possible immediate rewards, but only on the rewards that follow the specified action– so we do not need to take into consideration the term $\sum_a \pi(s,a)$ since it is only computed after the action was taken.

However, in the vast majority of situations the agents do not know *a priori* which actions are the best in order to make the optimal performance – the agent does not know the dynamics of the MDP - so he needs to discover them by a trial-and-error search, similar to what happens in instrumental learning. These estimation methods are called *Monte Carlo methods,* as they involve averaging over many random samples of actual returns. Algorithms facing this problem usually make use of a variable called prediction error (PE), defined as the difference between the expected and the actual reinforcement. Examples of these TD-algorithms are the SARSA, the Q-Learning and the Actor-Critic. Next we are going to explore each of them with more detail.

2.3.2 SARSA (On-Policy) and Q-Learning (Off-Policy) algorithms

In the SARSA (**S**tate-**A**ction-**R**eward-**S**tate-**A**ction) algorithm, state-value functions are updated in the following way:

$$\hat{Q}^{\pi}(s_t, a_t) \leftarrow \hat{Q}^{\pi}(s_t, a_t) + \alpha[r_{t+1} + \gamma \hat{Q}^{\pi}(s_{t+1}, a_{t+1}) - \hat{Q}^{\pi}(s_t, a_t)]$$
$$= \hat{Q}^{\pi}(s_t, a_t) + \alpha \delta_t \quad (2.9)$$

Here, $\hat{Q}^{\pi}(s_t, a_t)$ is the estimate of $Q^{\pi}(s, a)$, α denotes the learning rate ($0 \leq \alpha \leq 1$) and δ_t the prediction error. It follows that the TD error is based on the difference between the old action-value (what was expected) and the new action value plus the reward (what actually happened).

This method is called on-policy because it takes into account the next chosen action even though that action is not the best possible. In order to consider the second option, δ_t can be defined in a slightly different way, resulting in an off-policy method like the Q-Leaning:

$$\delta_t = r_{t+1} + \gamma max_a \hat{Q}^{\pi}(s_{t+1}, a_{t+1}) \tag{2.10}$$

If the proper conditions on the learning rate are assured and all the state-action pairs are visited infinitely often, both methods (SARSA and Q-Learning) will converge to the true optimal state-action values (in the latter case) or policy-dependent (in the former case).

After computing the state-action values the agent can simply choose the action with the highest action value – a greedy approach – or apply a more sophisticated rule, the *Softmax rule*, where actions are taken at a frequency proportional to their action-value. This allows for some randomness in the choices, and unvisited states can be chosen even though they may have lower state-action values:

$$Pr(a_t = a | s_t = s) = \pi_t(s, a) = \frac{e^{\beta Q(s,a)}}{\sum_{b \in A(s)} e^{\beta Q(s,b)}} \tag{2.11}$$

Beta is the inverse of temperature ($\beta \geq 0$), a trade-off between exploration-exploitation. In an extreme scenario, if $\beta = 0$ the choice is totally random (prevalence of the exploration effect), while its increase leads to the adoption by the agent of a more exploitation (and greedy) approach.

2.3.3 Actor-Critic algorithm

The Actor-Critic approach explicitly separates the policy evaluation from the value function [20]. By having a separate memory structure, the critic acts like an evaluator, receiving the state s_t and the reward r_t from the environment, and then computing a TD error based on them. This temporal difference prediction error is computed by the critic which will update the state value predictions $V(S)$, and then it is used by the actor to update its policy $\pi(S, a)$, evaluating the action just selected.

$$\hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha[r_{t+1} + \gamma \hat{V}^{\pi}(s_{t+1}) - \hat{V}^{\pi}(s_t)] \tag{2.12}$$

$$\delta_t = r_{t+1} + \gamma \hat{V}^{\pi}(s_{t+1}) - \hat{V}^{\pi}(s_t) \tag{2.13}$$

Here $\hat{V}^{\pi}(s_t)$ is the estimate of $V^{\pi}(s_t)$. Then the actor computes the policy:

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \eta[r_{t+1} + \gamma \hat{V}^{\pi}(s_{t+1}) - \hat{V}^{\pi}(s_t)] \tag{2.14}$$

Where η is the actor's learning rate. A positive error signal from the critic reinforces the critic to take the same action again, in that particular state, whereas a negative error signal inhibits such behavior.

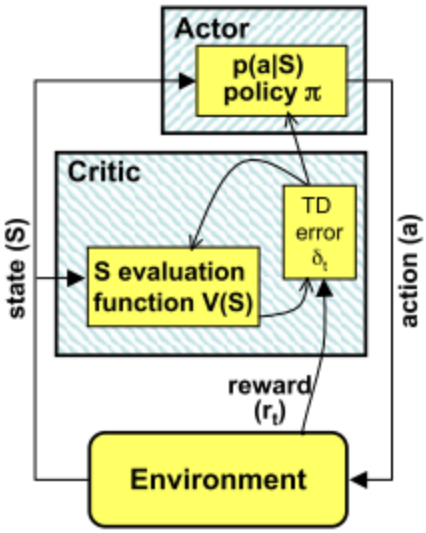


Figure 2.8: The actor-critic architecture (from [46]).

2.4 Synergy between Psychology, Neuroscience and Computer Science

Multiple lines of evidence have been linking the reinforcement learning framework to the function of dopaminergic neurons in the mammalian midbrain (by extracellular recordings during behavior tasks) [5, 6, 47] and, more recently, to data from human making-decision imaging experiments (fMRI) [7, 8, 41, 48]. A key link that arose is that dopamine appears to be correlated with a key RL signal, the temporal difference reward prediction error.

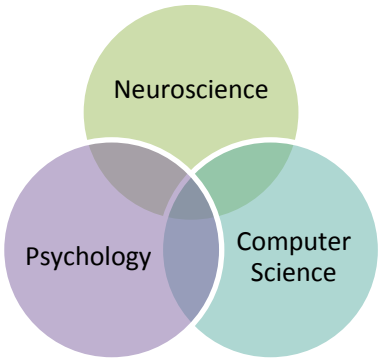


Figure 2.9: Synergy between Psychology, Neuroscience and Computer Science.

One of the most influential studies linking these three fields (Psychology, Neuroscience and Computer Science) was done by Wolfram Schultz [6] with his recordings from monkeys' midbrain, done while they were performing conditioning tasks. His conclusions demystified the idea that "dopamine equals to reward" showing that when learning was completed, dopaminergic cells stop responding, whereas monkeys start to show conditioned responses of anticipatory licking and arm movement – supporting that learning only happens when the reward is surprising or unexpected. This idea immediately gained attention in the scientific community, and several studies (e.g. [12]) found an analogy between the phasic firing of dopamine and the temporal difference reward prediction error. Soon, Montague et al. proposed a theoretical framework, the reward prediction error hypothesis of dopamine [11].

An illustration of this framework can be seen in the figure 2.10 from [10], where the recordings by Schultz [5] are compared with the Actor-Critic model:

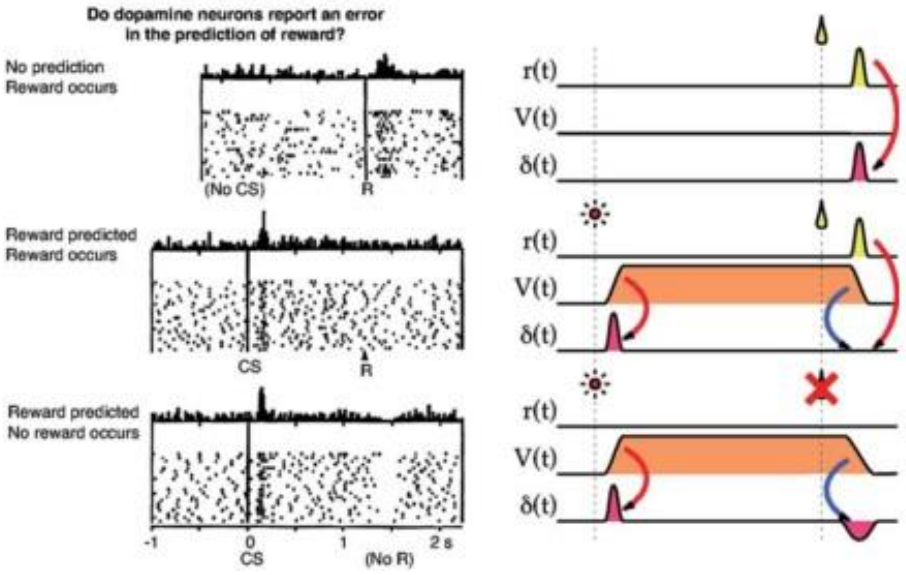


Figure 2.10: Dopamine neurons and reinforcement learning (from [9]).

- The recordings at the top of the figure were made before training, when there is no CS, and dopaminergic neurons fire in response to the unpredictable reward. Since the accumulative predicted reward V is zero for all states, the TD signal δ is equal to the reward signal r.
- After conditioning (recordings in the middle of the figure), a burst of dopamine following the occurrence of the CS (that predicts the reward) occurs and the TD signal δ is now positive, even if the reward has not yet been given.
- At the bottom, the predicted reward is omitted even though the CS is present - the firing of dopamine neurons showed a precisely-timed pause in firing, below their standard background firing rate. The TD signal δ is negative because reality was worse than expected. A similar comparison was done by Niv and colleagues [49] also showing error prediction back-propagation within the trial, during the learning process.

Additional experiments also showed that the theory can explain more complex aspects of learning, for example: i) phasic dopaminergic response is proportional to the magnitude and/or probability of the expected rewards, when they are of different magnitudes or of a probabilistic nature [50]; and ii) dopaminergic activity to a cue predicting a delayed is attenuated in proportion to the delay – with longer predicted delays eliciting a smaller dopaminergic response with line of the temporal discount parameter [47].

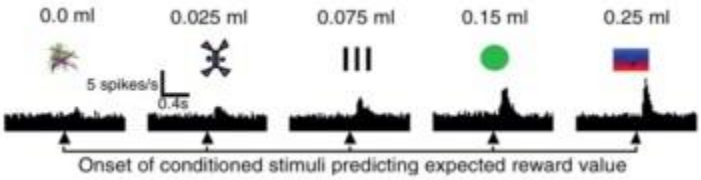


Figure 2.11: Phasic responses to a cue predicting reward are proportional to the magnitude of the predicted reward. (Adapted from [50]).

Another interesting point is that dopaminergic neurons do not seem to be involved in the signaling or in prediction errors for aversive outcomes ([51, 52]) even though they do appear to be connected with the absence of appetitive outcomes ([53]). This has led to the suggestion that dorsal raphe serotonin complements dopaminergic function in aversive learning [28].

However, the question, which method is implemented by the human brain is still under intense debate. At one hand, electrophysiological recordings (e.g. [47]) suggest a Q-learning framework. At the other hand, some studies (e.g. [20]) point to towards Actor-Critic approach suggesting that dopaminergic projections from VTA[1] (**V**entral **T**egmental **A**rea), which targets the ventral striatum and other limbic areas, might be responsible for calculating the state values (Critic), while that from SNc to dorsal striatum allows the policy learning (Actor).

This has also been supported with fMRI experiments where, using model-driven analysis, the reward prediction error in passive prediction-learning tasks (in which anticipation of rewards is dependent on the state but not on the agent's action – critic's role) showed a higher correlation with ventral striatal activity, whereas in active choice tasks (where reward's prediction is also based on action-values – actor's policy learning), the correlation was with both ventral and dorsolateral striatum ([54]). But whether BOLD (**B**lood **O**xygen **L**evel **D**ependent), signal can actually reflect the same underlying neural process in all brain areas is questionable. For instance, it is known that dopamine can directly affect the dilatation and contraction of local blood vessels ([55, 56]).

[1] Origin of the dopaminergic cell bodies

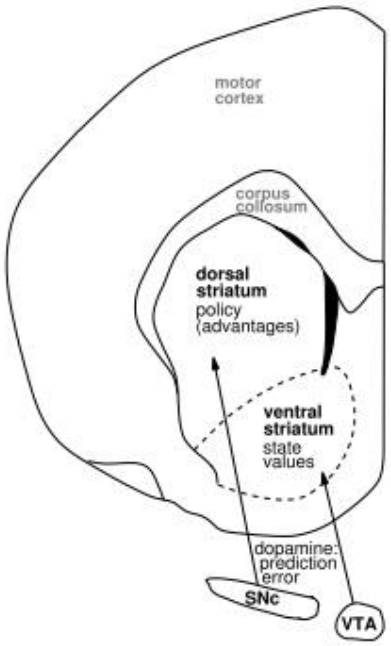


Figure 2.12: Suggested relationship for the actor-critic model in midbrain and striatum (from [57]).

Some studies propose that multiple reinforcement learning systems are implemented in the brain. In this case, there is a shift-mechanism between “model-free” RL methods, in which action selection is easier but much training is needed in order to predict reasonably well the future, and “model-based” methods that are more accurate and more rapidly adjustable face to changes in the environment, but more computationally demanding in terms of time and neural resources ([58]). At a neuronal level, separate cortico-basal-ganglia loops were implicated in each of these methods ([59]). However, how the brain actually makes that shift (if it exists) is far from being understood. One candidate for this function, from studies using rats, is the infralimbic cortex, another subarea of the medial prefrontal cortex ([60]).

So, with several arguments pointing in different directions is of utmost importance to try new approaches that can give us a better insight regarding this connection between RL and human decision-making.

3. Methods

Contents

As described before in the previous chapter, prediction errors in RL learning algorithms can be computed using different approaches. As a result, understanding which one better mimics the functioning of the brain is a hotly debated research topic. In this thesis, the design of a new task intended to contribute to this discussion is described.

3.1 Experimental Task

The proposed task, inspired by other probabilistic reinforcement Go/NoGo tasks, explores the key difference on how prediction errors are calculated: if using state-values (as the Actor-Critic), or if using action-values values (like in SARSA and Q-learning methods).

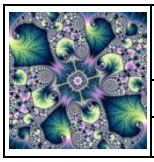
It is divided into five phases: it begins with two learning phases, with different trial types representing different conditions; a test phase (third phase); finally the fourth and the fifth phases analyze the use of subliminal images shown during the learning phases. A more detailed description of each phase, and the reasoning behind them, will be discussed further in the thesis at hand.

On the first three phases, four different conditions are used: three conditions differ in their **action** (pressing/not pressing a button) by **valence** (win or lose) interaction, while the forth represents a **neutral** condition, meaning that regardless the action performed, the outcome is always zero (a 0 appears on the screen). The trials that do differ in terms of action-valence interactions are named as follows: **go to win**, since the subject has to press the button to win a reward (+1 appears on the screen); **go to avoid losing**, where the subject should press the button to avoid a punishment (-1 appears on the screen); and **nogo to avoid losing**, because in order to avoid a punishment the subject needs to withhold from pressing the button. Although a fifth condition (nogo to win) would be needed to fully orthogonalize the action/valence in a 2 (reward/punishment) x 2 (Go/NoGo) design, it was not implemented on this task for reasons to be latter explained.

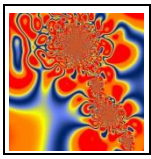
The outcome of each action, in a given trial, is probabilistic, so that the same action for the same type of condition can lead to different results. For instance, in the go to win condition, pressing the button can either lead to a reward (+1) with 80% probability; or to a neutral outcome (0) with 20% probability. The opposite action has similar opposite outcomes: if the subject withholds from pressing, the probability of receiving a neutral outcome is 80%, while receiving a reward has a probability of only 20%.

The probability distribution of outcome for all conditions is presented below, in figure 3.1:

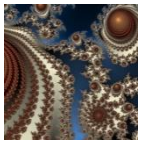
	go to win		
	-1	0	+1
Go	-	20%	80%
NoGo	-	80%	20%



	go to avoid losing		
	-1	0	+1
Go	-	100%	-
NoGo	80%	20%	-



	nogo to avoid losing		
	-1	0	+1
Go	80%	20%	-
NoGo	20%	80%	-



	neutral		
	-1	0	+1
Go	-	100%	-
NoGo	-	100%	-

Figure 3.1: Probability distribution of the outcomes for the go to win, go to avoid losing, nogo to avoid losing and neutral conditions. The possible outcomes are -1 (punishment), 0 (neutral), and +1 (reward). Images correspond to some of the fractals shown throughout the task.

The trials' outcome is stochastic in order to make the learning process more difficult, ensuring that the learning process is not too fast which allow us to better capture this effect of interest and that subjects keep their attention. Also, it should not be too difficult otherwise participants would just give up.

Additionally, the trials (states) are independent from each other, meaning that the transition probabilities do not depend on the performed action. For this reason, and remembering the formulation exposed in section 2.3, the discount rate is zero ($\gamma = 0$): the agent/subject does not have to take into account future rewards when choosing the best action to perform, for a given state/stimulus. Consequently, since ($\gamma = 0$), it is not possible to differentiate between SARSA and Q-learning models as they now share the same equation for the action-value iteration:

$$\hat{Q}^{\pi}(s_t, a_t) \leftarrow \hat{Q}^{\pi}(s_t, a_t) + \alpha[r_{t+1} - \hat{Q}^{\pi}(s_t, a_t)] \quad (3.1)$$

Keeping this in mind, from now on we will call this simplification a Q-learning class and, even though the task can't distinguish between the two models (SARSA and Q-learning), it does not affect our main goal. We intend to differentiate whether prediction errors are computed based on the state-value function or based on the action-value function and so we can still compare this Q-learning class and the AC model.

Although 4 conditions types were described, the one of interest to the problem under analysis is the **go to avoid losing**. Due to the way prediction errors are computed, the Q-value for the action go (Q-learning model) and the policy for pressing the button (AC model) assume different values for this trial type:

- In the **Q-learning model**, the Q value of each action is updated using the previous Q-value and the reward. Since every time a subject presses the button the reward is 0, the Q-value for the Go action does not change and thus remains 0 throughout the task (see eq. 3.1). On

the other hand, the Q-value for the NoGo action quickly becomes negative. Therefore, the subject chooses the Go action not because it has a positive Q-value, but because the Q value for the NoGo action is negative, and thus lower than the Go action.

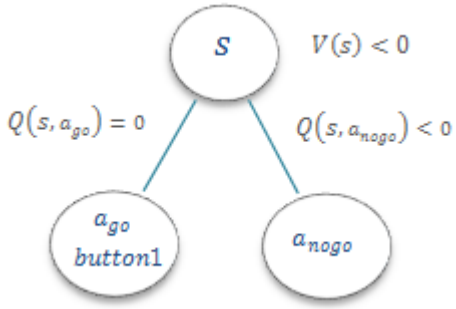


Figure 3.2: Diagram representing the updating of Q-values.

- In the **Actor-Critic** approach, the preference of each action is computed using the previous preference for that action and the prediction error (that takes into account the stimulus value $V(s)$).

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \eta\, \delta_t \qquad (3.2)$$

$$where\ \delta_t = r_{t+1} - \hat{V}^{\pi}(s_t)$$

With practice the stimulus value will tend to be negative and since the outcome is neutral for the Go action, a positive prediction error is computed whenever the button is pressed. Therefore the preference associated with the Go action will be positive. So, the subject chooses the Go action because it has a positive preference, and thus its preference will become higher than the preference for other actions in response to that stimulus.

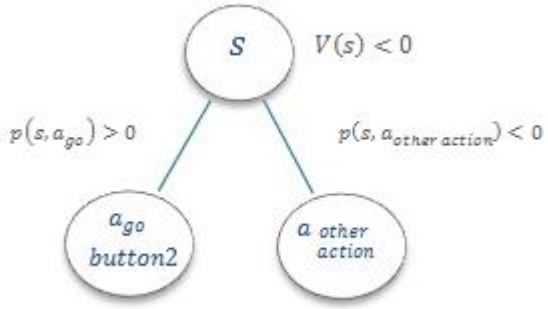


Figure 3.3: Diagram representing the updating of preferences.

As mentioned before, this condition is the only one where this discrepancy between the predictions of the two models occurs:

	go to win	go to avoid losing	nono to avoid losing	neutral
Q-learning	$Q(s, ago) > 0$	$\boldsymbol{Q(s, ago) = 0}$	$Q(s, ago) < 0$	$Q(s, ago) = 0$
Actor-Critic	$p(s, ago) > 0$	$\boldsymbol{p(s, ago) > 0}$	$p(s, ago) < 0$	$p(s, ago) = 0$

Table 3.1: Predictions of the 2 models. Simplified relation between Q-values and preferences for the 4 conditions.

All trials (occurring on the first two phases) follow a similar structure, presented in figure 3.4: first a blank screen occupies the screen for 500 ms as a preparation for the trial; then, a fixation point appears for 1000 ms; afterwards this the stimulus image (always a fractal image) is displayed with a maximum duration of 1500 ms – during this time the subject is allowed to press the button (performing a Go trial) or withhold from pressing the button (executing a NoGo trial). The rest of the sequence depends on whether the trial is a Go trial or a NoGo trial.

- In a Go trial, immediately after pressing the button, another fractal image is presented but only for a very short period of time (16 ms) – it is a subliminal image – and followed by the same stimulus image but now with a transparency layer (confirming the subject that the button press was valid) for 500 ms. After the image removal, there is a loading phase during which a progress indicator appears in order to give the participant a sense of time (3000 ms) before the feedback is made available for 1000 ms. The feedback can be a reward (a green +1), a punishment (a red -1) or a neutral outcome (black 0).
  §
- In a NoGo trial, the sequence is similar, but the stimulus image stays on the screen for 1500 ms and it is immediately followed by the loading period.

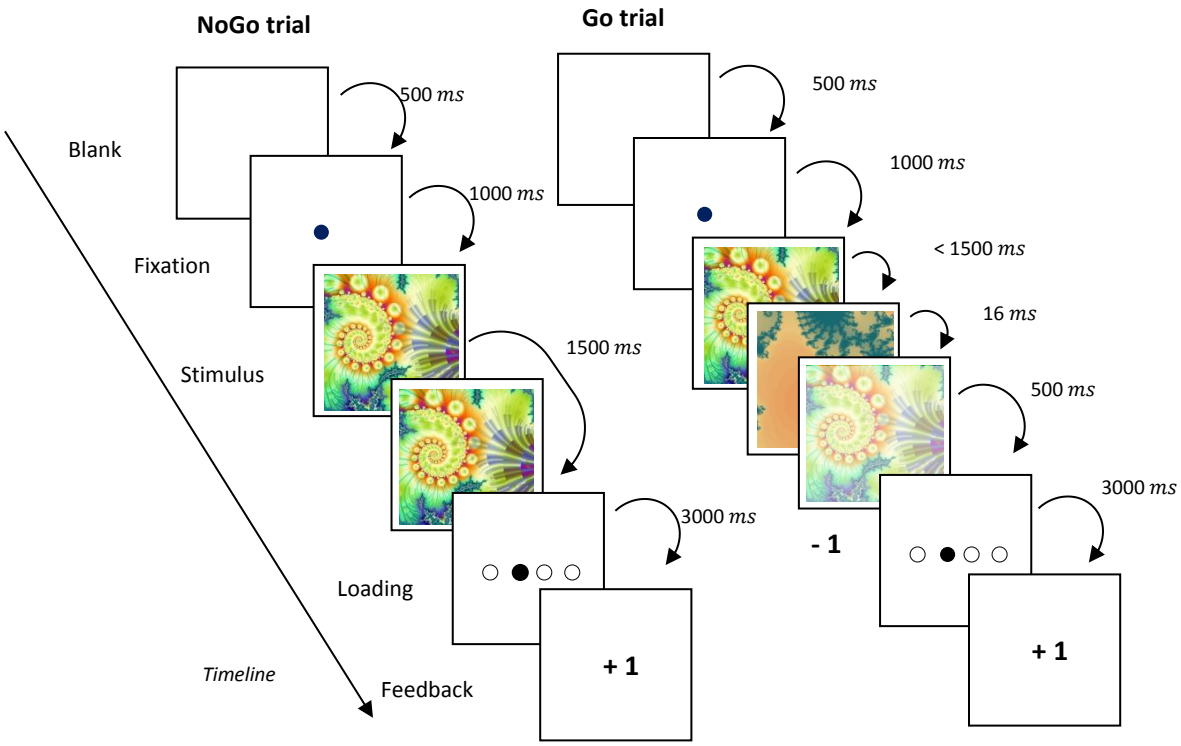


Figure 3.4: Experimental paradigm during the learning phases.

Having presented the designed trials and their structure, we will now discuss each phase in detail:

- **Learning phases (1st and 2nd Phases)**

1st Phase

In this phase four independent images are presented, each with a 25% probability: one representing a **go to win** condition (*Distractor +, phase 1*), other the **nogo to avoid losing** condition (*Distractor -, phase 1*), and two different images both acting as a **neutral** condition (called *Neutral and Go to avoid losing*. The *Go to avoid losing* image is named as such because it will be associated with a **go to avoid losing** condition on the 2nd phase, despite being associated with a neutral trial in this phase).

The phase does not have a fixed duration: it finishes only when the subject performs 10 Go trials in each of the two images acting as neutral condition. This leads to some subjects performing the phase faster than others. To avoid prolonging the phase too much (for example, in cases where the subject never presses the button) a guard was included: after 100 trials the task ends automatically.

In this phase subjects have to press a button, designated as button 1.

2nd Phase

Again, there are four different image trials. Two images are replaced for new images (*Distractor +, phase 2*; *Distractor -, phase 2*), representing the **go to win** condition and the **nogo to avoid losing** condition, respectively. The other two are kept constant as in the first phase: the *Neutral*, which continues to act as a **neutral** condition (the outcome is always zero independently of the action performed); and the *Go to avoid losing* image, now associated with a **go to avoid losing** condition, instead of a neutral condition. Here, in a Go trial the feedback is always zero, but in a NoGo trial there is 80% probability of receiving a punishment (-1) and 20% probability of getting a neutral outcome (0).

The duration of the phase is again dynamic: the subject needs to press the button 10 times in the *Neutral* and in the *Go to avoid losing* images.

Another difference versus the previous phase is the change of the button pressed by the subject. Now subjects need to press a button, called button 2 – this change will actually play an important role in the test phase.

For the 1st and 2nd phases, the update of both Q-values (Q-learning) and preferences (Actor-Critic model) for the Go to avoid losing condition, can be predicted as follows:

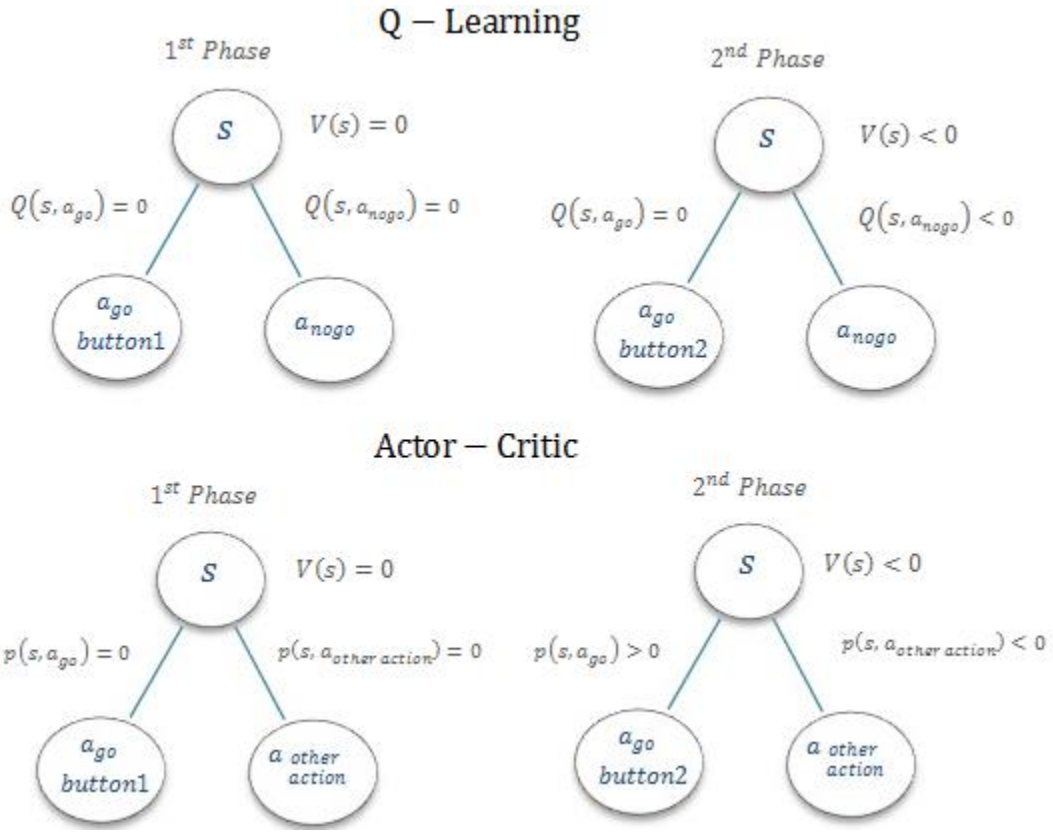


Figure 3.5: Diagram showing the differences of both methods for the Go to avoid losing condition.

- **Test Phase (3rd Phase)**

During the test phase, the subjects are presented with all the stimulus images of both learning phases. Now the option is no longer to press (or not) a button but instead which button to press: either button 1 (the one used in the 1st phase) or button 2 (the one used in the 2nd phase). Since it is a test phase, subjects need to remember what they have just learnt and have to apply it; no feedback is given now (there are no rewards, punishments nor neutral feedback). A schematic diagram of a trial type is shown in fig 3.6.

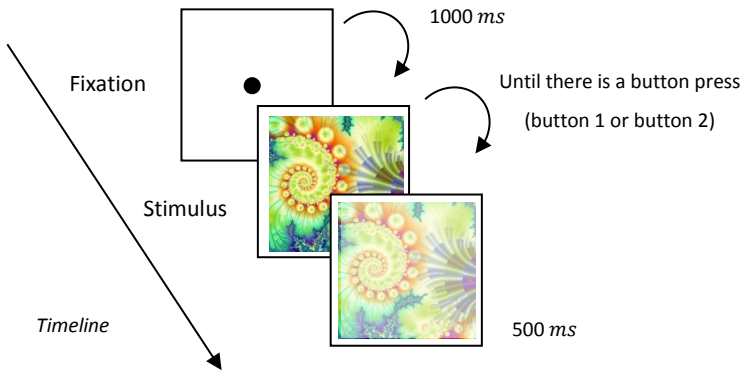


Figure 3.6: Experimental paradigm during the Test Phase.

In phase 3 each image is presented a fixed number of times:

- All the distractors (*Distractor +, phase 1*; *Distractor -, phase 1*; *Distractor +, phase 2*; *Distractor -, phase 2*) are shown 5 times each;
- The *Go to avoid losing* image and the *Neutral* image are displayed 10 times each.

Due to the difference regarding the update of Q-values and preferences, there are two possible expected outcomes for the **go to avoid losing** condition:

1) If the **Q-learning model** is the one being implemented by the human brain, at the end of both learning phases (1st Phase and 2nd Phase) the Q-value for the Go action is zero (and obviously higher than the negative Q-value for the NoGo action). Since the subject associated a Q-value for the Go action equal to zero in the first two phases, he should choose indifferently between both buttons during the test phase (3rd phase);

2) If an **Actor-Critic** approach is followed, the subject associates a higher preference value for button 2. Then, it is expected that, when presented with the image and obliged to make a choice, he will prefer to press button 2.

For all the other conditions the expected behavior is the same regardless of the model implied. We anticipate that the subject will press button 1 in *Distractor +, phase 1's* image and button 2 in *Distractor +, phase 2's* image because he associated a positive Q-value/preference for the Go action, for these images, to button 1 (in the 1st Phase) and to button 2 (in the 2nd Phase), respectively. Concerning the negative distractors, since the subject associated a negative Q-value/preference for the Go action, for those images, he will tend to press the button of the alternative phase. In other words, we predict that the subject will more likely press button 1 in *Distractor -, phase 2's* image and button 2 in *Distractor -, phase 1's* image.

For the **neutral** condition there is no expected tendency. Since the outcome was always zero throughout the whole task, this image will be a tool to measure the subject's bias to press either button and to better conclude if the other conditions were learned.

It is worth to mention that the task design was a complex and thoughtful exercise. The goal is always to minimize the possible *confounders*, as it is essential to control all the extraneous variables that may influence the results (the only variable that should influence the results is the one being studied). During the design of the task, 3 major sources of confounders were studied, and their minimization had an impact on the task's design:

- **Association effect –** Learning is an iterative process done every time the subject performs an action. Since we will analyze performance during the test phase we need to make sure

that the subject had the same number of Go trials, concerning the images that we want to compare during both learning phases (the *Go to avoid losing* and the *Neutral* images)– meaning that he learnt the action equally well in the two phases. Therefore, we need to make sure that in each phase, the button is pressed the same number of times.

Solution: Both learning phases only finish when the subject pressed the button 10 times for the *Go to avoid losing* and the *Neutral* images. This was the reason why the condition **nogo to win** was not used. Even though this condition, like the **Go to avoid losing** condition, would provide different results in the test phase (in this situation, $Q(s, ago) = 0$ and $p(s, ago) < 0$), the imposition of a fixed number of Go trials is unreasonable: the majority of learners would learn that he should avoid pressing in that condition, never reaching the obligatory number of Go trials.

- **Serial position effect** - The response tendency, either due to association or preference, might also depend on the serial position of each action. In other words, the subject could press button 2 not because he exhibits a preference for it but because it happened later on in the task - this event is called recency effect. The opposite effect could also happen, the so-called primacy effect, and the subject presses button 1 because it was first he learnt. This problem could be overcome by counterbalancing the 2 phases. However, in this case, that is not possible because the Go to avoid losing image must act as a neutral condition (for staying with a null value), before acting as a Go to avoid losing condition (and therefore acquiring a value different from 0). Otherwise, undesired state-values and action-values would be carried to the next phase.

  Solution: A possible way for dealing with this problem is taking into account the subject's behavior to the Neutral image during the 3rd Phase. It shows us if a subject is more biased to press a certain button.

- **Image/outcome association** - Tendency to associate certain images with a determined outcome (e.g. negative outcome with the color red)

  Solution: Images and conditions were counterbalanced across subjects. The same image trial is not always associated with the same condition or even with the same phase.

Between each phase, subjects need to read the instructions explaining how the phase works (its rules) and that the decision to press a button (in the 1st and 2nd Phase) or which button to press (in the 3rd Phase) should be made as fast as possible. This element of speed plays an important role, since we are interested in model-free learning. This procedure is a way to isolate this type of learning and to avoid the possibility of any model-based construction (the subject is not supposed to mentally discover the task's algorithm).

3.2 Safety signal theory: The use of subliminal images

As explicitly shown in figure 3.4 during both learning phases, every time the subject chooses to perform a go action (Go trial), a subliminal image is presented, immediately after pressing a button and before the trial image reappears (with a transparency layer). It is important to note that the type of subliminal images, the duration of them and when they appeared was also a thoughtful process and several tests were need before achieving the final task version. The use of subliminal images is another possible tool to support the definition of which model (Q-learning or Actor Critic model) is being used during the brain's learning process.

Contrarily to the learning process described so far (stimulus-action-reward), which is clearly a case of instrumental conditioning, the use of subliminal images in this task leads to a classical conditioning learning process (stimulus-reward). Despite the change between instrumental and classical models, Q-values or preferences and state-value predictions $V(S)$ are also computed.

Since every stimulus image has a subliminal image associated to it, the subject will update a state-value $V(S)_{subliminal}$ to each subliminal image across the task. Recalling equation 2.12:

$$\hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha[r_{t+1} + \gamma\hat{V}^{\pi}(s_{t+1}) - \hat{V}^{\pi}(s_t)] \tag{2.12}$$

$$since\ \gamma = 0 \Longrightarrow \hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)] \tag{3.3}$$

And due to the temporal sequence between the trial image and the associated subliminal image (the trial image always precedes the subliminal image), in this classical conditioning arrangement, the outcome is computed as a prediction error that can be positive, neutral or negative. So, concerning the subliminal image associated with to the *Go to avoid losing* image, we will have the following equation for updating $V(S)_{subliminal}$:

$$\hat{V}^{\pi}(s_t)_{subliminal} \leftarrow \hat{V}^{\pi}(s_t)_{subliminal} + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)_{subliminal}] \tag{3.4}$$

$$r_{t+1} = r_{t+1_{go}} - what\ is\ expected \tag{3.5}$$

According to which model is being used, the variable $'what\ is\ expected'$ can be either $\hat{Q}^{\pi}(s_t, a_{t,go})_{trial}$ or $\hat{V}^{\pi}(s_t)_{trial}$. As outlined before, this is the key difference between the two models.

Therefore, even though the value-states of all subliminal images start at zero, they evolve in one of two ways:

1) If the **Q-learning** model is the one being used, the valence of the subliminal images associated with the negative distractors (*Distractor -, phase 1*; *Distractor -, phase 2*) should be lower than the one associated with the images of the neutral and go to avoid losing conditions (in both phases) and those should be lower than the valence of the subliminal images correspondent to the positive distractors (*Distractor +, phase 1*; *Distractor +, phase 2*).

   According to this model, the state-value of the subliminal image associated with the *Go to avoid losing* image will always be zero until the end of the second phase – since $\hat{Q}^{\pi}(s_t, a_{t,go})_{trial} = 0$.

   So, no difference regarding the valence of the subliminal images of the *Neutral* compared to the valence of the *Go to avoid* losing images (in both learning phases) is predictable.

2) If it is an **Actor-Critic** approach that is being used, the order of the valence of the subliminal images obtained for each of the 4 conditions will be slightly different. Lower valences are expected to be associated with the negative distractors, followed by the ones associated with the *Neutral* images of both phases and with the *Go to avoid losing*'s image of the 1st Phase (remembering that even though it is named Go to avoid losing, this image acts as a neutral condition in the 1st Phase). It then follows the valence associated with the Go to avoid losing image of the second phase, and finally, valence correspondent to the positive distractors will have the highest values.

   The state-value of the subliminal image associated with the **go to avoid losing** condition will acquire a positive value:

$$\hat{V}^{\pi}(s_t)_{subliminal} \leftarrow \hat{V}^{\pi}(s_t)_{subliminal} + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)_{subliminal}] \quad (3.6)$$

$$r_{t+1} = r_{t+1_{go}} - \hat{V}^{\pi}(s_t)_{trial} \quad (3.7)$$

   The reward for this condition, in a Go trial, is always zero ($r_{t+1_{go}} = 0$), and with time $\hat{V}^{\pi}(s_t)_{trial} < 0$, making $\hat{V}^{\pi}(s_t)_{subliminal} \rightarrow \hat{V}^{\pi}(s_t)_{subliminal} > 0$.

   In fact, we can relate this reward to a positive prediction error because, even though the actual outcome is always zero the subject was expecting an aversive outcome that turned out to be neutral –acting as a safety signal [61, 62]

Since the two most important valences will be the ones given to the subliminal images associated with the **go to avoid losing** condition (from the 2nd Phase) and to the **neutral** condition (from that same phase), across all the subjects, we made sure that those conditions were

associated with each fractal subliminal image the same number of times. For a given subliminal image (for instance, fractal 5), if it appears 4 times associated with the go to avoid losing condition (so in 4 different tests), it will also appear 4 times with the neutral condition.

- **Subliminal Perception Phase (4th Phase)**

In order to guarantee that the subliminal images were in fact subliminal, a *perceptual discrimination task* was used as a control for awareness at the end of the Test Phase. Here, sixteen different images were shown: 8 of them were the subliminal images displayed through the 1st and the 2nd Phases, and the other 8 had never appeared before (control stimuli). The display order of the images was counterbalanced across subjects.

Subjects were asked if they had or had not seen each image before. The response was given manually, by pressing one of two keys: **S** (the initial of **yes** in Portuguese), in case of "I have seen the image"; or **N** (the initial of **no** in Portuguese), in the case of "I don't have seen the image".

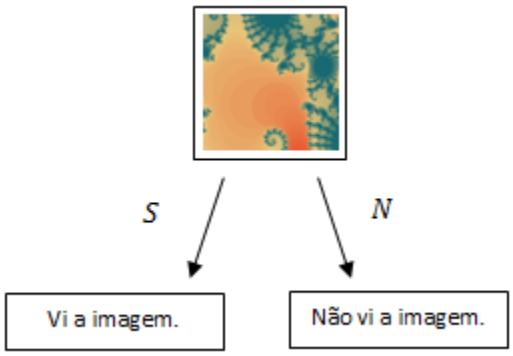


Figure 3.7: Experimental paradigm during the 4th Phase: **S** if the subject saw the image before ("Vi a imagem" is "I have seen the image" in Portuguese; or **N** if the subject did not see the image before ("Não vi a imagem" is "I do not have seen the image").

Each subject has his own **criterion**, meaning that a person may only choose the affirmative option if he is 100% sure that she had seen it before, while another may choose it even having doubts about it – they may have a different *bias*. This criterion is also a trade-off used to minimaze false positives and false negatives. The response is also inherently noisy: affected by external noise and internal noise inherent to neuronal responses (the same image can produce different neuronal activities in different times, under the same conditions). We can then think that after seeing an image an *internal response*, associated with a neural activity, emerges.

If we associate a hypothetical internal response curve distribution to the group of subliminal images and another one to the group of control images, the criterion threshold will divide the graph into four sections: hits, misses, false alarms, and correct rejections:

- For both hits and false alarms, the internal response is greater than the criterion – and so the subject answers “yes” – however in the former case the image actually appeared before, and in the latter case it did not.
- In cases of misses the subject answers “no” to an image that was shown subliminally, and in correct rejections trials the answer is also “no” but in this case that is the correct answer (since the image has never been shown).

It is clear that, depending on the applied criterion, a subject can more easily say “yes” (increasing the number of hits and false alarms) or “no” (increasing the number of misses and correct rejections).

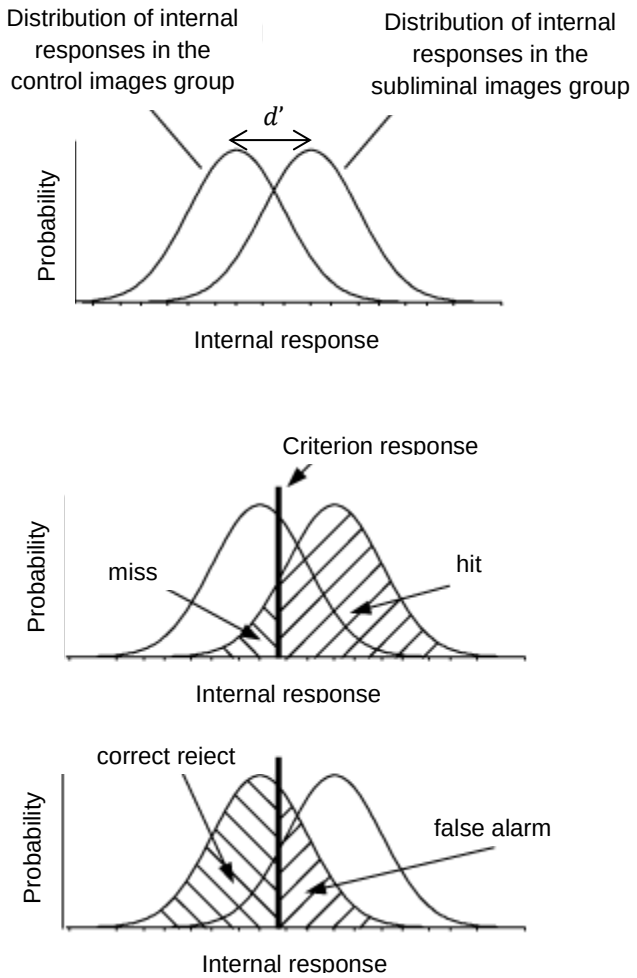


Figure 3.8: Reasoning behind the d’ calculation.

Thus, the discriminability of the images depends both on the separation and the spread of the two distributions. This can be formulated by the sensitivity index (d') [63]. Since we calculate a d'

value for each subject, the standard deviation (giving the spread) for both curves will be assumed to be 1, simplifying the calculations. So, d' will be the distance between the means of the curves, translated by the difference between normalized rates of hits and false alarms (z-values) (eq. 3.9). Therefore, a d' close to zero can be interpreted as a lack of conscious access during the main task.

	Responded Absent ("Não vi a imagem")	**Responded Present** ("Vi a imagem")
Stimulus Present	Miss	Hit
Stimulus Absent	Correct Rejection	False Alarm

Table 3.2: The four categories: Miss, Correct Rejection, Hit and False Alarm.

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate) \quad (3.9)$$

- **Subliminal Valence Phase (5th Phase)**

In the 5th Phase, only the subliminal images were showed and again the order was counterbalanced across subjects. Subjects were forced to choose a number from 1 to 9 representing "*How much*" they like each image. This evaluation was used as being the valence of each subliminal image ($V(s)_{subliminal}$). Then, results were analysed in order to see which model (Q-learning or AC) best fits the data and so, better the brain's function.

It can be argued that classical conditioning was unconsciously processed, by showing that the presence of subliminal images exerts an indirect influence on participants behaviour (through their choices during this evaluation phase), but fail to reach awareness in the direct d' test.

4. Behavior analysis

Contents

4.1 Subjects

35 healthy adults (22 males and 13 females; age range 22-58 years; mean age 29.7±12.2) participated in the study, performing the task previously described. All subjects provided written informed consent for the experiment, which was approved by the local Ethics Committee for the Health Care of University of Lisbon.

4.2 Behavior analysis

The analysis of subjects' behavior will be presented by order: starting with the data from both Learning Phases, followed by the data from Test Phase, and at the end, the results from the Subliminal Perception Phase and from the Subliminal Valence Phase.

4.2.1 Learning Phases (1st and 2nd Phases)

Before analyzing the data that will give us insight regarding which model better fits human decision making behavior, it is important to check whether subjects actually learnt the proper associations between stimuli and responses during the 1st and during the 2nd Learning Phases.

In figure 4.1 and 4.2, the time probability across subjects of making the go action for each condition is depicted, in the 1st and in the 2nd Learning Phase, respectively. From the temporal dynamics of the curves, subjects appear to have learned correctly all four conditions:

- In the first phase, subjects correctly learnt to withhold from pressing in the *Distractor* – image and to press in the *Distractor* + image, clearly seen by the red and green curves, respectively (fig. 4.1). Concerning the other two images, both acting as a neutral condition, subjects decreased the pressing rate. This can be explained by a cost to perform the pressing action – since the reward was always zero independently of the action performed, the participant could opt not to press just because not reacting requires less effort. That effect might become more pronounced with time.

- In the second phase, both distractors images were well learnt (again, showed by the red and green curves; fig. 4.2). For the *Neutral* image (black curve) subjects decreased the pressing rate, similar to the previous phase. Finally, the blue curve, representing the *Go to avoid losing* image, shows that subjects seem to have learnt this condition as well, staying the Go probability always between 65 and 80%, on average.

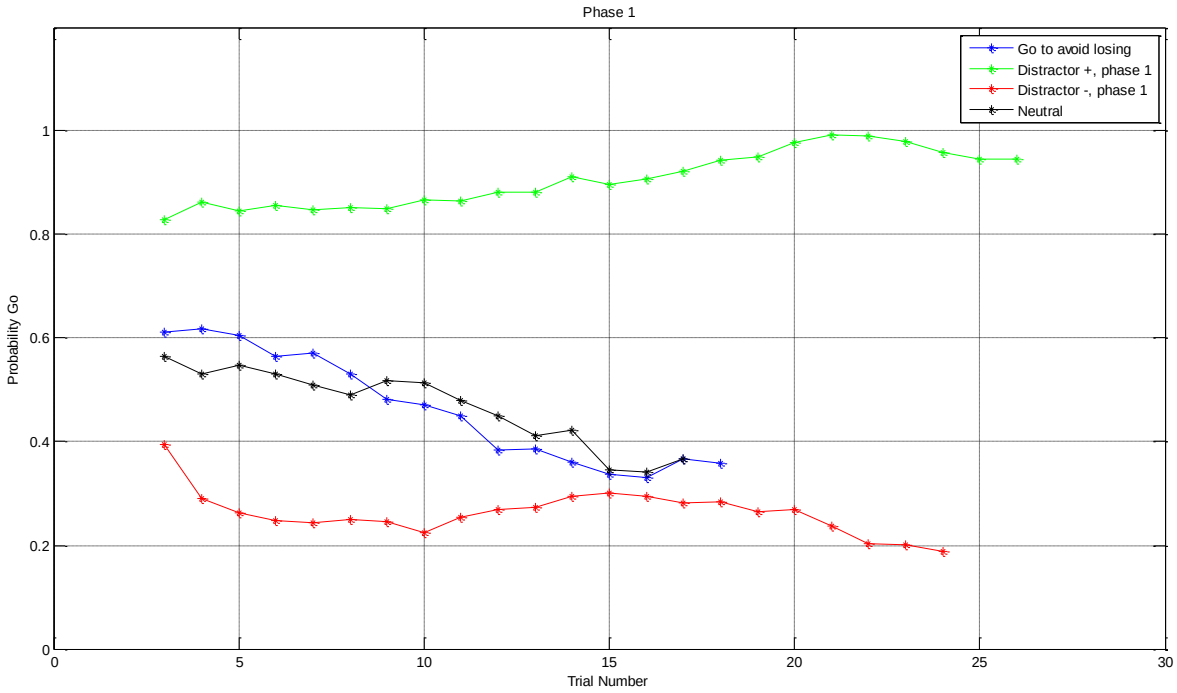


Figure 4.1: Time varying probabilities, across subjects of making a go response for each condition. Data were convolved with a central moving average filter with a length of 5 for the 1st phase. Data is only presented when there are at least 16 subjects in a trial.

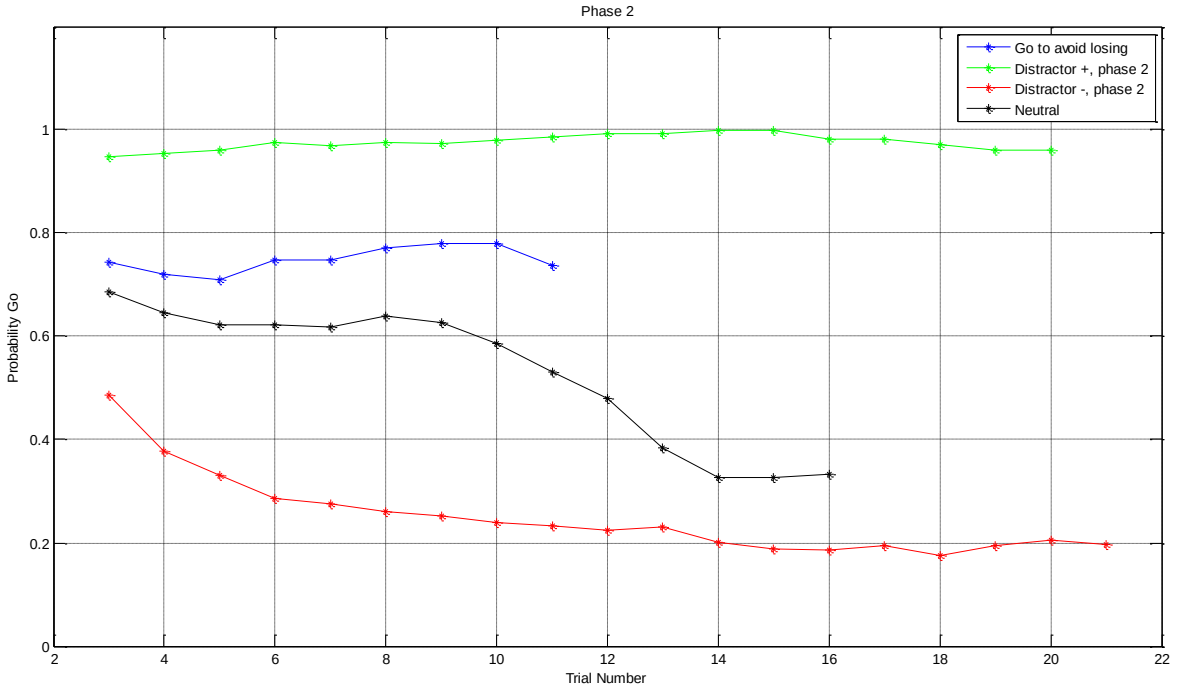


Figure 4.2: Time varying probabilities, across subjects of making a go response for each condition. Data were convolved with a central moving average filter with a length of 5 for the 2nd phase. Data is only presented when there are at least 16 subjects in a trial.

Note that, since in each phase, subjects needed to perform the Go action 10 times in the **neutral** and in the **go to avoid losing** condition (the reasoning behind this imposition was explained in the previous chapter), the number of trials per phase is not a fixed. So, it is normal that different conditions may have different number of trials.

Even though such descriptive look to the learning curves seems to confirm the expected learning, a statistical analysis was performed. For that, each phase was divided in two blocks (1st and 2nd half). Then, the difference between the probability of doing the Go action for each condition in the 2nd and 1st half was calculated. The difference was used to perform a one-way ANOVA (**AN**alysis **O**f **VA**riance) with factor of condition, in each phase.

- In the first phase, the ANOVA showed a main effect of condition ($F(3,102) = 3.556, p = 0.017$), which indicates that there is a statistically significant difference, at least between two conditions. However, a post hoc paired t-test demonstrated that between the *Go to avoid losing* and the *Neutral* image there was not a statistically significant learning difference ($p = 0.172$). This was already expected since both images were acting as a **neutral** condition in this phase.

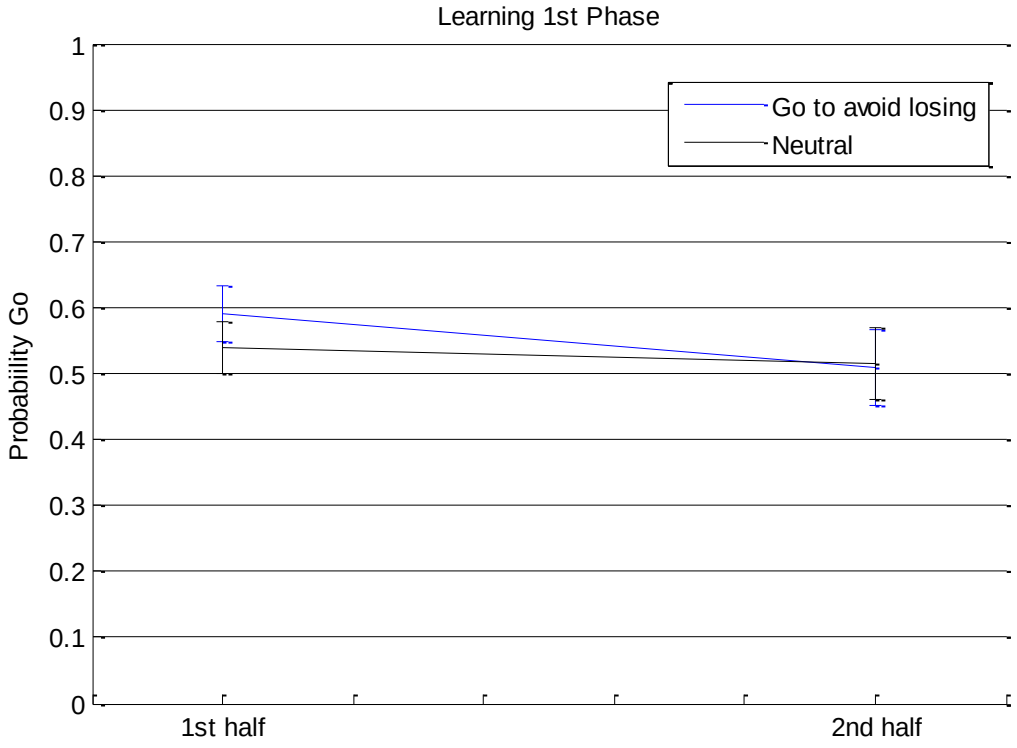


Figure 1.3: Probability of Go in the 1st and 2nd half for the go to avoid losing and neutral conditions, during the first learning phase. Error bars represent the standard error of the mean (S.E.M.).

- In the second phase, the ANOVA showed a main effect of condition ($F(3,102) = 3.556, p < 0.001$). But now, when applying a post hoc paired t-test between the two aforementioned conditions, there was a statistically significant learning difference ($p = 0.033$), demonstrating that the **Go to avoid losing** condition was well learnt and it differentiated from the **Neutral** condition.

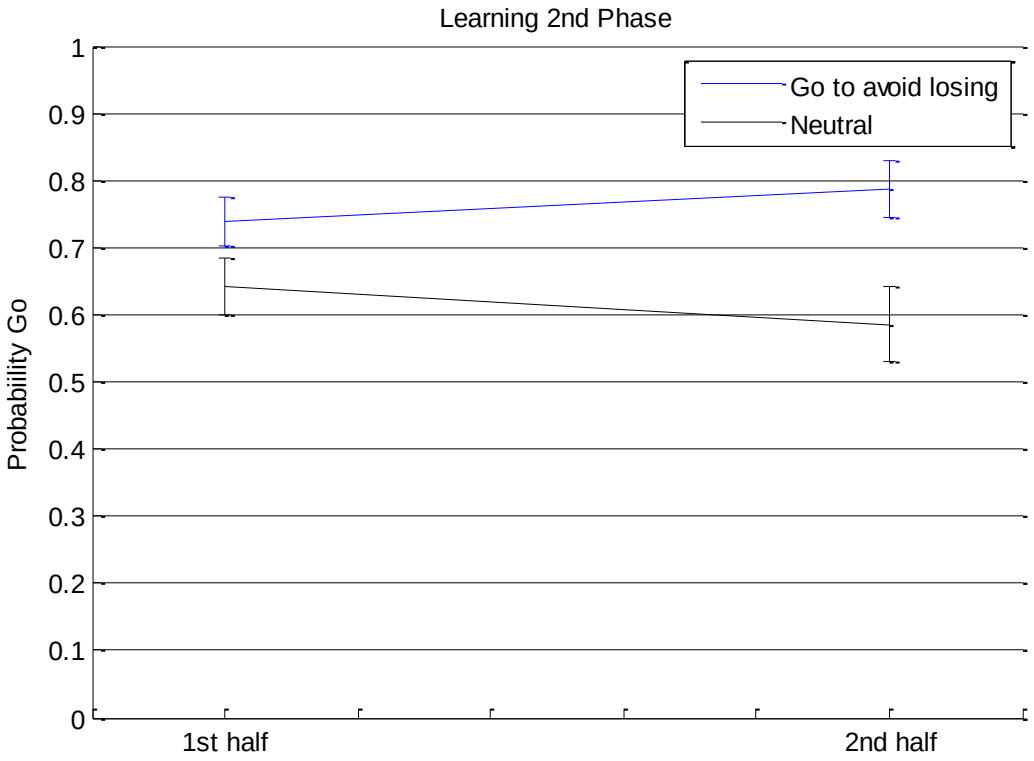


Figure 4.4: Probability of Go in the 1st and 2nd half for the go to avoid losing and neutral conditions, during the second learning phase. Error bars represent the standard error of the mean (S.E.M.).

4.2.2 Test Phase (3rd Phase)

In the Test Phase subjects were forced to choose between button 1 (first phase's button) or button 2 (second phase's button).

In figure 4.5, the percentage of button 2 choices for each trial image is depicted.

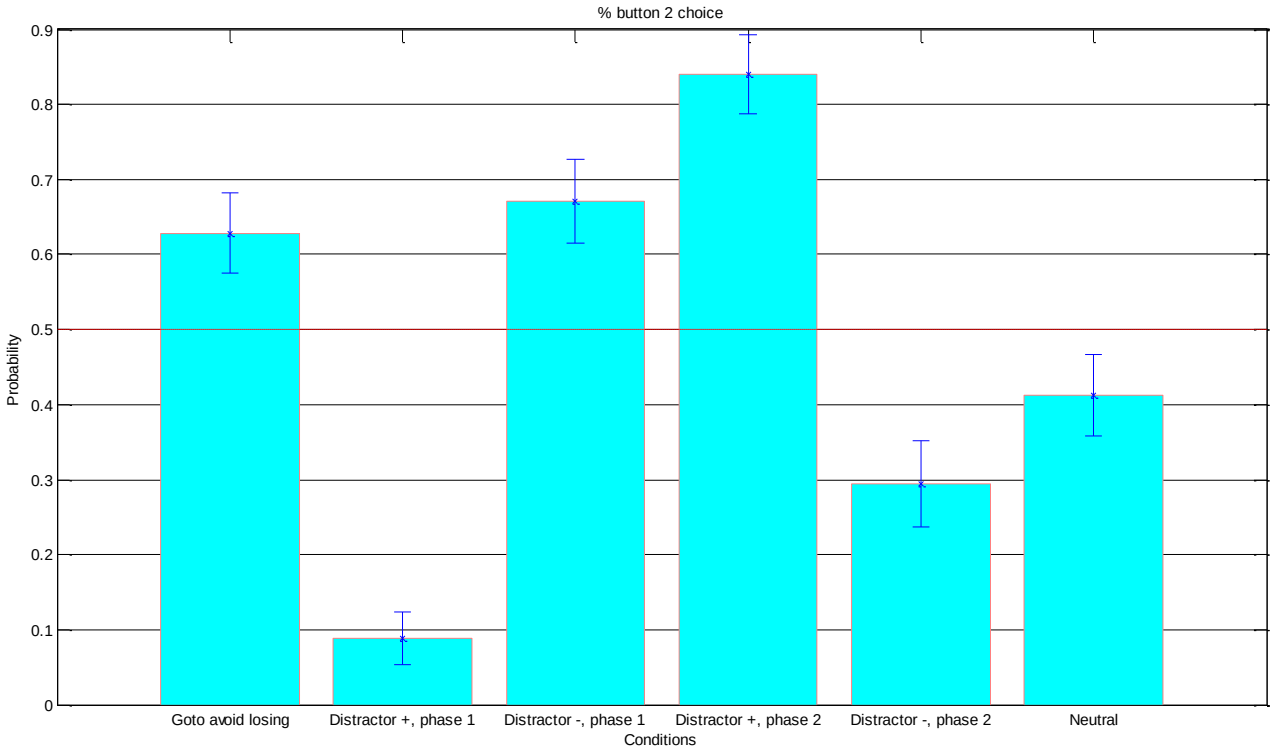


Figure 4.5: Percentage of button 2 choice in the 3rd phase for all the conditions. Error bars represent the standard error of the mean (S.E.M.)

With regard to the analysis of the distractors images, it can be seen that whenever a positive preference was associated with a certain button during the learning phases, in the Test Phase participants tended to press that same button (for the associated image). I In other words, since the *Distractor +, phase 1* image was associated with a positive preference in the 1st Phase, participants were likely to press button 1 in the 3rd Phase (Test Phase) ($\mathbf{91,14 \pm 3,5}\%$) rather than button 2 (only $\mathbf{8,88 \pm 3,5}\%$). The same analogy can be made for the *Distractor +, phase 2* image but in this case the positive preference is associated with button 2, resulting in a preference for this button in the Test Phase ($\mathbf{84 \pm 5,30}\%$) against the few presses on button 1 (only $\mathbf{16 \pm 5,30}\%$).

If, contrarily, participants learnt that the correct action was to withhold from pressing (leading to a negative preference to press) when an image appeared in one of the learning phases, subjects preferred to press the button used in the other learning phase. In more detail, subjects preferred to press button 2 ($\mathbf{67,14 \pm 5,58}\%$) in response to the *Distractor -, phase 1* image and button 1 ($\mathbf{70,57 \pm 5,42}\%$) to *Distractor -, phase 2* image during the 3rd Phase.

However, the behavior just described does not help us to answer which type of reinforcement learning method is being used since the results until now are in accordance with both of them – the previous explanation could also have been made with Q-values instead of preferences. The crucial

result is the subject's choice in the *Go to avoid losing* image (in the Test Phase) taking into account the *Neutral* image's response as a bias. As explained previously, a preferential tendency to press button 2 after seeing the *Go to avoid losing* image in the 3rd Phase is in line with the Actor Critic model, whereas a lack of preference when choosing the button to press is in consonance with the Q-learning framework.

Due to the possible serial position effect, the performance in the *Neutral* image is a critical aspect. For example, one could argue that a higher percentage of button 2 choice in the **go to avoid losing** condition was due to the subject's preference for the button learnt latter and not due to a higher computed preference – so, this bias needs to be taken into consideration. However, subjects have chosen almost equally the two buttons ($\mathbf{58,74 \pm 5},42\%$ the button 1 and $\mathbf{41,26 \pm 5},42\%$ the button 2) when the stimulus image was the *Neutral*. As already mentioned, the lower value might be explained by a cost to perform the pressing action.

So, the tendency to press button 2 in the *Go to avoid losing* image ($\mathbf{62,86 \pm 5},35\%$) was subtracted by the bias to press button 2 (given by the percentage of button 2 choices in the *Neutral*) for each subject. Then, the mean of those subtractions and the standard error of the mean (S.E.M) were calculated, being $21{,}59\%$ and $6{,}57\%$, respectively.

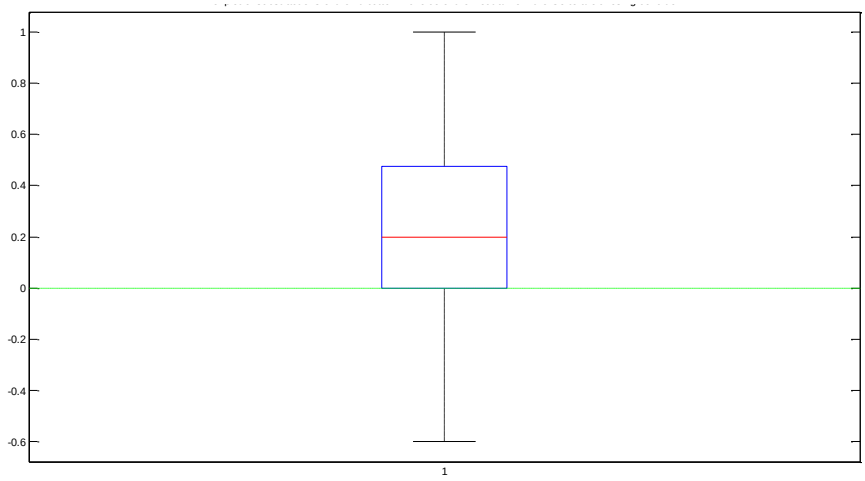


Figure 4.6: Boxplot of subtractions of the proportion of button 2 choices of the Neutral from the Go to avoid losing condition.

Figure 4.6 exhibits the boxplot in relation to the subtractions of the proportion of button 2 choices of the 35 subjects in the **neutral** condition from the **go to avoid losing** condition, in the 3rd phase. The red line represents the median (approximately 0.20, $\mathbf{20}\%$); the extremes of the blue rectangular the 25th and the 75th percentiles (showing that 75% of the subjects actually pressed more often the button 2 in the **go to avoid losing** condition than in the **neutral** condition); and the whiskers extend to the most extreme data points, which the algorithm considers to not be outliers – outliers would be plotted with individual asterisks, if they were present.

Additionally, a one-way ANOVA with factor of condition, in the data with the percentages of button 2 choice, showed a strong main effect of condition ($\mathbf{F(5, 170) = 25.345, p < 0.001}$). With regard to the two most important conditions (the **go to avoid losing** and the **neutral**), a post hoc one-tail paired t-test revealed that there is a significant difference between them ($\mathbf{p < 0.001}$), showing that the percentage of button 2 choices in the **go to avoid losing** condition is significantly higher than the percentage of button 2 choices in the **neutral** condition.

In sum, all the listed arguments strongly suggest that subjects performed this task accordingly to the Actor-Critic model.

4.2.3 Subliminal Perception Phase (4th Phase)

From an overall view of the subjects' responses, it can be easily perceived that the majority declare they have not seen any of the subliminal images shown. In fact, from the 35 participants, 30 gave a negative response to all of the 16 images presented during the perceptual discrimination task. From the others 5 subjects, 2 reported to have seen the same number of images that were actually shown across the task and of images that were used in this phase just as a control. The remaining 3 subjects gave a positive response more frequently to images that were used as subliminal images. A summary of all 35 responses is reported bellow.

	Responded Absent ("Não vi a imagem")	**Responded Present** ("Vi a imagem")
Stimulus Present	95,71%	4,29%
Stimulus Absent	97,86%	2,14%

Table 4.1: Proportion of responses, in the 4th Phase, for each type of stimulus (images).

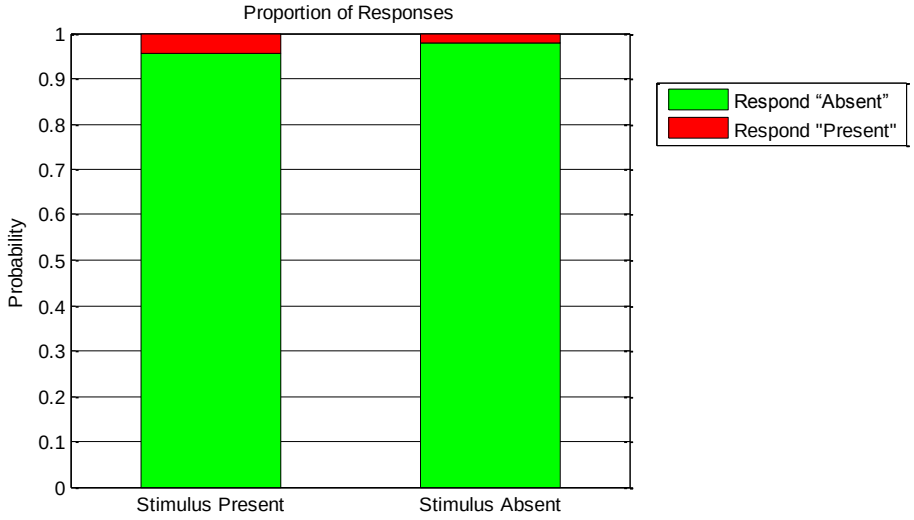


Figure 4.7: Proportion of responses about whether subliminal perception occurred.

The analysis just presented suggests that subliminal images were in fact subliminal, but an objective measurement reinforced it. The sensitivity index (d') was calculated (equation 3.9) for each subject based on the responses during the 4th phase. Then, it was demonstrated that this measure was not significant from zero using a one-tailed paired t-test ($p = 0.0506$). This value was close to be significant but a boxplot of the d' values showed that the three subjects that stated to have seen, more frequently, the subliminal images could be considered as outliers. When excluding them, all d' values would be zero ($p = 1$).

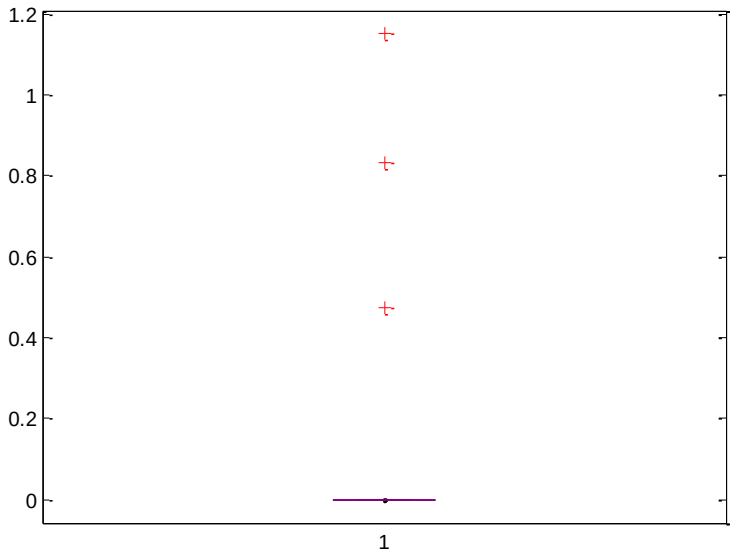


Figure 4.8: Boxplot of all d' values. The three outliers are depicted as crosses.

4.2.4 Subliminal Valence Phase (5th Phase)

The average of all given valences ratings, per image, is presented in figure 4.9. A growing tendency can be seen: the negative distractors' subliminal images were associated with lower values, followed by the Neutral's subliminal images, then the Go to avoid losing's subliminal image from the 2nd Phase and finally, with the highest value, the subliminal image' valence correspondent to the positive distractor of the 2nd Phase. Remember that the Go to avoid losing image from the 1st Phase acts as a neutral condition; this might explain the small value associated with its subliminal image.

This tendency is consistent with the Actor Critic model and so, with the results from the Test Phase that were already discussed.

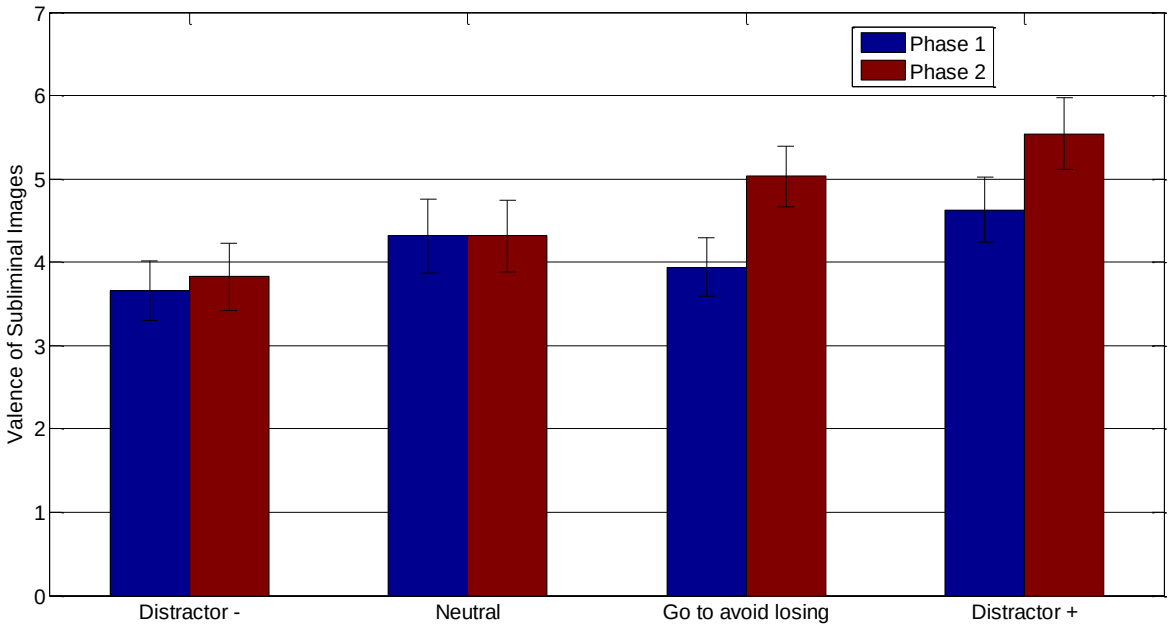


Figure 4.9: Subliminal images' valence. Error bars represent the standard error of the mean (S.E.M.).

However, we need to keep in mind that different people might use the value-range in a different way: a subject may only choose values between 1 and 3, while another may use the whole range available (from 1 to 9). So, it is important to normalize the valence's values to their z-scores, for each subject before averaging them out. This normalization describes where a value is located in the individual distributions - for instance, a negative z-score means that the original score was below the mean of the subject– while also scaling them by his standard deviation.

$$z_{subject,image} = \frac{x_{subject,image} - \mu_{subject}}{\sigma_{subject}} \quad (4.5)$$

Figure 4.10 presents the subliminal images' valences after the z-score normalization. The results are not surprising: negative distractors' subliminal images values are negative, followed by the Neutral's subliminal images values close to zero (to the mean), then the Go to avoid losing's subliminal image from the 2nd Phase acquires a positive value, as well as the positive distractor of the 2nd Phase. Again, the value of the Go to avoid losing's subliminal image from the 1st Phase is lower than what was expected, but remember that its stimulus image associated was acting as a neutral condition.

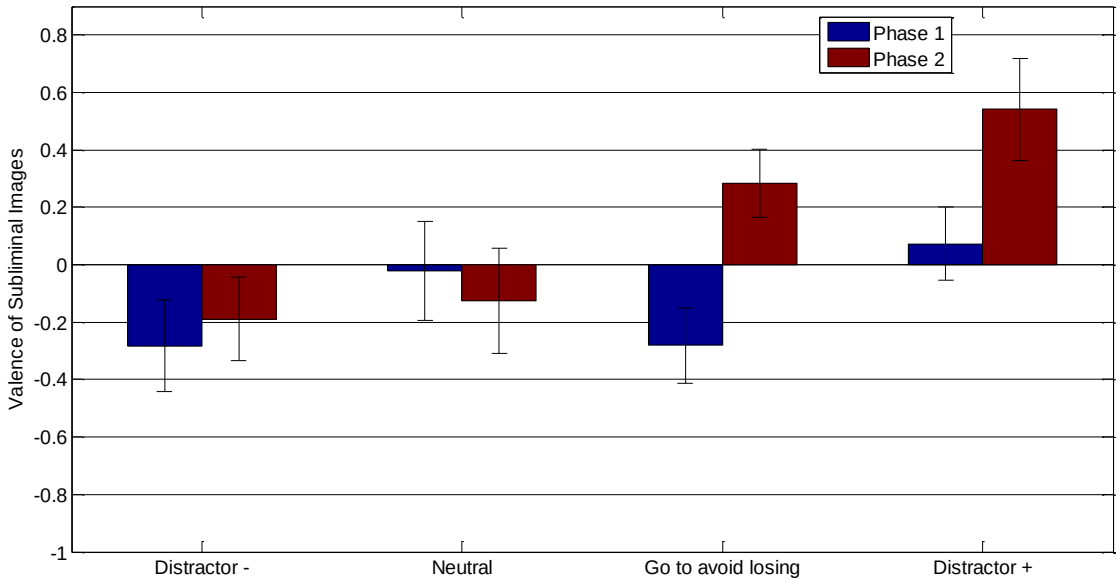


Figure 4.10: Subliminal images' valences normalized by the z-scores. Error bars represent the standard error of the mean (S.E.M.).

A one way ANOVA with factor of condition was also performed using the mean valences normalized and it indicated the existence of a main effect of condition ($F(7,238) = 3.121, p = 0.004$).

Since the most important aspect is, as it has been explained throughout the whole study, the difference of values between the *Go to avoid losing* and the *Neutral* images of the 2nd Phase - we are interested in determining whether the subliminal image associated with the *Go to avoid losing* from the 2nd Phase assimilated a value significantly higher than the value of the *Neutral*'s subliminal image of the 2nd Phase – a post hoc one-tailed paired t-test was done. It revealed that there is a significant difference between them ($p = 0.0418$), strongly pointing towards the Actor Critic model.

Again, doing this using the z-scores, instead of the actual given values, only scales those differences by $\frac{1}{\sigma_{\text{subject}}}$.

$$z_{subject,Go\ to\ avoid\ losing2} = \frac{x_{subject,Go\ to\ avoid\ losing2} - \mu_{subject}}{\sigma_{subject}} \quad (4.6)$$

$$z_{subjectNeutral2} = \frac{x_{subject,Neutral2} - \mu_{subject}}{\sigma_{subject}} \quad (4.7)$$

$$z_{subject,Go\ to\ avoid\ losing2} - z_{subjectNeutral2} =$$

$$\frac{x_{subject,Go\ to\ avoid\ losing2} - \mu_{subject}}{\sigma_{subject}} - \frac{x_{subject,Neutral2} - \mu_{subject}}{\sigma_{subject}} =$$

$$\frac{x_{subject,Go\ to\ avoid\ losing2} - \mu_{subject} - x_{subject,Neutral2} + \mu_{subject}}{\sigma_{subject}} =$$

$$\frac{x_{subject,Go\ to\ avoid\ losing2} - x_{subject,Neutral2}}{\sigma_{subject}} \quad (4.8)$$

Finally, another argument in favor of the Actor-Critic framework is the fact that the *Go to avoid losing* image of the 2nd Phase acquired a z-score valence significantly positive, while the *Neutral* image did not. A t-test of the z-score of each condition against zero resulted in $p = 0.0229$ and $p = 0.4969$, respectively).

5. Conclusions and future work

Contents

In this last chapter, final comments about the present study are presented, as well as ideas for future improvements.

5.1 Conclusions

The aim of this study was to investigate whether prediction errors in the humans' brain are determined by state values (Actor-Critic model) or by action values (Q-learning framework). To achieve that a new Go/NoGo task was specifically designed in order to address this question and tested in 35 healthy subjects. Then, the behavioral data was analyzed.

Before focusing on the analysis concerning our main question, it was important to verify if the subjects actually learnt the task, specially the go to avoid losing condition compared with the reference neutral condition. The descriptive analysis seemed to confirm the desired learning and an inferential statistical analysis showed that the go to avoid losing condition differentiated from the neutral one, in terms of a higher probability of Go trials (confirming that learning occurred). This is an essential finding, especially in a new task, since we need to make sure that the inherent difficulty and duration of the task were appropriate. Several pre-versions of the task were needed before reaching this point. For example versions using conditions with different probability distribution (making the task harder or easier) or with different stimulus' duration.

Since in this study, classical conditioning was used through the use of subliminal images, it was also important to verify that those images were in fact subliminal. For that the sensitivity index d' was calculated for every participant. Once again, pre-versions of the task had different durations for the subliminal images, as well as, different chosen images, in order to make sure that the purpose was being well accomplished. The results of these pre-versions were not in the focus of this thesis and are not reported here.

To answer the central question, two approaches were used: one based on the instrumental conditioning and another on classical conditioning (subliminal images):

- On the one hand, the subjects' behavior in the Test Phase supports that all conditions were well learnt and showed that there was a significant preference to choose the button from the 2nd Phase (button 2) when the stimulus image was the **go to avoid losing** condition, taking into account the **neutral**. This supports the Actor-Critic model ($p(s, ago)_{button\,1} = 0$; $p(s, ago)_{button\,2} > 0$) whereas a balanced behavior would point to the Q-learning framework ($Q(s, ago)_{button\,1} = 0; Q(s, ago)_{button\,2} = 0$).

- On the other hand, regarding the subliminal images, subjects were asked to give them a valence between 1 and 9, an indirect measurement of their state-values ($\mathrm{V}(S)_{\mathrm{subliminal}}$). It was also revealed that the **go to avoid losing image**'s valence was

significantly higher than the **neutral** image's valence, of the 2nd Phase. In fact it was also shown that the former image acquired a z-score valence significantly positive while the latter image did not. Again, this supports the Actor-Critic model since with time, for the **go to avoid losing**'s subliminal image $\widehat{V}^{\pi}(s_t)_{subliminal} \rightarrow \hat{V}^{\pi}(s_t)_{subliminal} > 0$, while for the **neutral**'s it remains zero ($\widehat{V}^{\pi}(s_t)_{subliminal} = 0$). If the Q-learning had been the model employed, both state-values would have kept the neutral value. This change underlies the fact that depending on the model being used prediction errors are calculated using state-values and action-values, respectively.

In brief, the behavioral data's analyzes from the two approaches were in agreement, and strongly pointing towards the Actor-Critic model.

This was in accordance with several studies that defend that different neuronal areas may be responsible to calculate prediction errors and state values (acting like the Critic), while others may use them to learn an action-selection policy (being the Actor) [19, 57]. fMRI studies also support this theory [6-8, 54] and so, our findings.

Additionally, if we look to the basal ganglia Go/NoGo (BG-GNG) model, only the Actor-Critic model (2.6) is biologically congruent:

i. Thinking about the Q-learning model: in the beginning of the learning process, for the Go to avoid losing condition the prediction error is negative (when the outcome is negative), and so a dopamine dip would occur. The basal ganglia neuroanatomical model suggests that the Go pathway (direct pathway) would be weakened while the NoGo pathway (indirect pathway) will be strengthened.
ii. On the other hand, in the Actor-Critic approach the choice is not made between doing an action (go action) and not doing it (no go action), but instead doing that action (e.g. pressing a button) or any other action (that can be as diverse as scratching his nose or standing up). Giving the separate memory structure, it is possible to update the value-function every time that condition is presented (regardless of the performed action) whereas the preference for the go action (its policy) is only updated when the subject chooses to perform the action. So, in the go to avoid losing condition the state-value becomes to be negative leading to a positive prediction error when the action is performed: this strengthens the Go pathway and diminishes the strength of the NoGo pathway, the desired consequence.

Concluding, our findings suggest that healthy humans follow the Actor-Critic model, basing their decisions on prediction errors computed with state-values. This is in agreement with several studies [6-8, 19, 54, 57] and consistent with the basal ganglia Go/NoGo (BG-GNG) model proposed. The new task design greatly contributed to reveal this.

5.2 Future work

Even though the behavioral investigation done allows us to make properly motivated arguments supporting that the human brain follows the Actor-Critic model, the potential scope of analysis done with data resulted from the programmed task is much wider. That is why it is so important to focus our attention in designing new tasks (or modifying existing ones).

For instance, the task's output also gives us the reaction time of each trial (the time between the stimulus' presentation and the subject's action) for the learning phases and for the test phase. It also gives us the reaction times in the 4^{th} and 5^{th} phases. A possible correlation could exist between the reaction times in the test phase and the subliminal images' valence: for example, we would expect that the higher the valence given in the Go to avoid losing's subliminal image of the 2^{nd} Phase is, the smaller the reaction times in the test phase for that same condition – since that condition should have been better learnt in that case.

Additionally, a model-based approach should be taken into consideration. Fitting the data with several models and comparing them would allow us to estimate numerous parameters, giving us a deeper knowledge about which method is used, and other parameters such as learning rates.

Although it is not possible to perform invasive electrophysiological recordings in humans, like it is done in animals, indirect measures could be done, like BOLD. So, performing the task inside an fMRI scanner and performing an fMRI model-based analysis could also give more information about how prediction errors are determined in the striatum.

Finally, the same task should be done in more subjects. Increasing the number of subjects (N) would allow us to have a better sample of the general population, reducing the variability and highlighting the differences under study. For example, in our particular case, increasing N could highlight the differences in the % of choosing button 2 in the Test Phase and dilute some of the unexpected results in the 5^{th} phase, like the low valence of the Distractor -, phase 1's image.

Bibliography

1. Domjan, M., *Principles of Learning and Behavior Active Learning*. 6 ed. 2010: Cengage Learning.
2. Sutton, R.S. and A.G. Barto, *Reinforcement Learning: an Introduction*. 1998: The MIT Press.
3. Kaelbling, L., M. Littman, and A. Moore, *Reinforcement Learning A Survey.* Journal of Artificial Intelligence Research, *arXiv preprint cs/9605103* (1996).
4. Schultz, J., *Predictive Reward Signal of Dopamine Neurons.* Journal of Neuroplysiology, 80.1 (1998): 1-27.
5. Schultz, W., *A Neural Substrate of Prediction and Reward.* Science, 1997. **275**(5306): p. 1593-1599.
6. Schultz, J., P. Apicella, and T. Ljungberg, *Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task.* The Journal of Neuroscience, 13.3 (1993): 900-913..
7. Roesch, M.R., et al., *Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain.* Eur J Neurosci, 2012. **35**(7): p. 1190-200.
8. O'Doherty, J.P., A. Hampton, and H. Kim, *Model-based fMRI and its application to reward learning and decision making.* Ann N Y Acad Sci, 2007. **1104**: p. 35-53.
9. Guitart-Masip, M., et al., *Go and no-go learning in reward and punishment: interactions between affect and effect.* Neuroimage, 2012. **62**(1): p. 154-66.
10. Doya, K., *Reinforcement learning: Computational theory and biological mechanisms.* HFSP Journal, 2007. **1**(1): p. 30-40.
11. Montague, P.R., P. Dayan, and T.J. Sejnowski, *A framework for mesencephalic dopamine systems based on predictive hebbian learning.* The Journal of Neuroscience, 1996. **16**: p. 1936-1947.
12. Maia, T.V. and M.J. Frank, *From reinforcement learning models to psychiatric and neurological disorders.* Nat Neurosci, 2011. **14**(2): p. 154-62.
13. Shah, A., *Psychological and Neuroscientific Connections with Reinforcement Learning*, in *Reinforcement Learning: State of the Art*, M. Wiering and M.v. Otterlo, Editors. 2012.
14. Frank, M.J., *Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Nonmedicated Parkinsonism.* Journal of Cognitive Neuroscience, 2005: p. 51-72.
15. Albin, R., A. Youg, and J. Penney, *The functional anatomy of basal ganglia disorders.* Trends in Neuroscience, 1989. 366-375.
16. DeLong, M.R., *Primate models of movement disorders of basal ganglia origin.* Trends in neurosciences, 13.7 (1990): 281-285.

17. Palminteri, S., et al., *Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes.* Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19179-84.
18. Worbe, Y., *Reinforcement Learning and Gilles de la Tourette Syndrome: Dissociation of Clinical Phenotypes and Pharmacological Treatments.* Archives of general psychiatry, 68.12 (2011): 1257-1266.
19. Szepesvári, C., *Reinforcement Learning Algorithms for MDPs.* 2009.
20. Barto, A.C., *Adaptive critic and the basal ganglia*, in *Models of information processing in the basal ganglia*. 1994, MIT Press: Department of Computer Science, University of Massachusetts.
21. Holland, P.C. and R.A. Rescorla, *The effect of two ways of devaluing the unconditioned stimulus after first-and second-order appetitive conditioning.* Journal of Experimental Psychology: Animal Behavior Processes, 1.4 (1975): 355.
22. Rescorla, R.A. and A.R. Wagner, *A theory of Pavlovian conditioning_ Variations in the effectiveness of reinforcement and nonreinforcement.* In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II. New York: Appleton-Century-Crofts, 1972.
23. Thorndike, E.L., *Animal inteligence: Experimental studies.* New York: Macmillan, 1911.
24. Maia, T.V., *Reinforcement learning, conditioning, and the brain: Successes and challenges.* Cogn Affect Behav Neurosci, 2009. **9**(4): p. 343-64.
25. Adams, C.D. and A. Dickinson, *Instrumental responding following reinforcer devaluation.* The Quarterly Journal of Experimental Psychology, 1981. **33**: p. 109-142.
26. Colwill, R.C. and R.A. Rescorla, *Postconditioning devaluation of a reinforcer affects instrumental responding.* Journal of experimental psychology: animal behavior processes, 11.1 (1985): 120.
27. Dayan, P. and Y. Niv, *Reinforcement learning: the good, the bad and the ugly.* Curr Opin Neurobiol, 2008. **18**(2): p. 185-96.
28. Daw, N.D., Y. Niv, and P. Dayan, *Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.* Nature neuroscience, 8.12 (2005): 1704-1711.
29. Holland, P.C. and M. Gallagher, *Amygdala circuitry in attentional and representational processes.* Trends in cognitive sciences, 3.2 (1999): 65-73.
30. Killcross, S. and P. Blundell, *Associative representations of emotionally significant outcomes.* Emotional Cognition: From brain to behaviour, 44 (2002): 35.
31. Yin, H.H., *The role of the dorsomedial striatum in instrumental conditioning.* European Journal of Neuroscience, 22.2 (2005): 513-523.
32. O'Reilly, R. and Y. Munakata, *Explorations on Computational Neuroscience*. The MIT Press
33. Dayan, P. and L.F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* The MIT press, 2001.

34. Ludvig, E.A., M.G. Bellemare, and K.G. Pearson, *A Primer on Reinforcement Learning in the Brain: Psychological, Computational, and Neural Perspectives*, in *A primer on reinforcement learning in the brain Psychological computational and neural perspectives*, E. Alonso and E. Mondragon, Editors.
35. Hebb, D.O., *The organization of behavior. .* New York: Wiley, 1949.
36. Reynolds, J.N., B.I. Hyland, and J.R. Wickens, *A cellular mechanism of reward-related learning.* Nature, 413.6851 (2001): 67-70.
37. Wickens, J.R., A.J. Begg, and G.W. Arbuthnott, *Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro.* Neuroscience, 70.1 (1996): 1-5.
38. Wickens, J.R., et al., *Dopaminergic mechanisms in actions and habits.* J Neurosci, 2007. **27**(31): p. 8181-3.
39. Mink, J., *The basal ganglia focused selection and inhibition of competing motor programs.* Progress in Neurobiology, (1996): 381-425.
40. Gerfen, C.R., *Molecular effects of dopamine on striatal-projection pathways.* Trends in Neuroscience, 23 (2000): S64-S70.
41. O'Doherty, J., et al., *Dissociable roles of ventral and dorsal striatum in instrumental conditioning.* Science, 2004. 304.5669 (2004): 452-454.
42. Hikida, T., et al., *Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior.* Neuron, 2010. **66**(6): p. 896-907.
43. Nambu, A., H. Tokuno, and M. Takada, *Functional significance of the cortico–subthalamo–pallidal 'hyperdirect'pathway.* Neuroscience research 43.2 (2002): 111-117.
44. Gerfen, C.R., *D1 and D2 dopamine receptor regulation of striatonigral and striatopallidal neurons.* The Neurosciences, 250.4986 (1990): 1429-1432..
45. Shen, W., et al., *Dichotomous dopaminergic control of striatal synaptic plasticity.* Science, 2008. **321**(5890): p. 848-51.
46. Niv, Y., *Reinforcement learning in the brain*, in *Psychology Department & Princeton Neuroscience Institute*1997, Princeton University
47. Roesch, M.R., D.J. Calu, and G. Schoenbaum, *Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. .* Nature neuroscience, 10.12 (2007): 1615-1624.
48. O'Doherty, J.P., et al., *Temporal difference models and reward-related learning in the human brain.* Neron, 2003. **28, 329–337**: p. 9.
49. Niv, Y., M.O. Duff, and P. Dayan, *Dopamine, uncertainty and TD learning.* Behav Brain Funct, 1.6 (2005): 1-9.
50. Tobler, P.N., C.D. Fiorillo, and J. Schultz, *Adaptive coding of reward value by dopamine neurons.* Science 307.5715 (2005): 1642-1645.

51. Mirenowicz, J. and W. Schultz, *Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli.* Nature 379.6564 (1996): 449-451.
52. Tobler, P.N., A. Dickinson, and W. Schultz, *Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm.* The Journal of neuroscience, 23.32 (2003): 10402-10410.
53. Bayer, H.M., B. Lau, and P.W. Glimcher, *Statistics of midbrain dopamine neuron spike trains in the awake primate.* Journal of Neurophysiology, 98.3 (2007): 1428-1439.
54. Daphna, J., Y. Niv, and E. Ruppin, *Actor–critic models of the basal ganglia: New anatomical and computational perspectives.* Neural networks 2002. **15**(4): p. 535-547.
55. Attwell, D. and C. Iadecola, *The neural basis of functional brain imaging signals.* Trends in neurosciences, 2002. **25**(12): p. 621-625.
56. Krimer, L.S., *Dopaminergic regulation of cerebral cortical microcirculation. APA* Nature neuroscience 1998: p. 286-289.
57. Daw, N., Y. Niv, and P. Dayan, *Actions, policies, values and the basal ganglia.* Recent breakthroughs in basal ganglia research 2005 (2006): 91-106.
58. Dayan, P., Y. Niv, and N. Daw, *Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.* Nature neuroscience, 8.12 (2005): 1704-1711.
59. Balleine, B.W., *Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits.* Physiology & behavior, 86.5 (2005): 717-730.
60. Killcross, S. and E. Coutureau, *Coordination of actions and habits in the medial prefrontal cortex of rats.* Cerebral Cortex 13.4 (2003): 400-408.
61. Tanimoto, H., M. Heisenberg, and B. Gerber, *Even timing turns punishment to rewards.* Nature, 2004. **430**: p. 983.
62. Seym0, et al., *Opponent appetitive-aversive neural processes underlie predictive learning of pain relief.* Nat Neurosci, 2005. **8**: p. 1234-1240.
63. Vermeiren, A. and A. Cleeremans, *The Validity of d' Measures.* Plos one, 2012.