# A behavioral investigation of the algorithms underlying reinforcement learning in humans

Ana Catarina dos Santos Farinha

*Under supervision of Tiago Vaz Maia*

Instituto Superior Técnico
Instituto de Medicina Molecular
Faculdade de Medicina da Universidade de Lisboa

November, 2014

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.

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

## 1. Introduction

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. The idea of decision making as a response to rewards and punishments has been supported with several animal behavioral studies, leading to the development of classical and instrumental conditioning theories.

In classical conditioning a behavior (conditioned response, CR) comes to be elicited by a stimulus (conditioned stimulus, CS) that has acquired its power through an association with a biologically significant stimulus (unconditioned stimulus, US) [1].

In instrumental conditioning the subject's behavior is adjusted accordingly to its consequences [1]. This interactive characteristic is the main difference between the two learning types and it is formulated by Thorndike's law [2] – when a response in presence of a stimulus is followed by a satisfying outcome, the stimulus-response (S-R) connection should be fortified. And, oppositely, when it is followed by an undesirable outcome, then the S-R connection should be weakened.

Considering that punishments can act as negative rewards, it is not surprising that artificial intelligence developers soon tried to create algorithms that, making use of the reward signal, allowed self-learning machines [1, 3]. This initiated an influential area called *reinforcement learning* (RL) from which temporal difference (TD) algorithms are the most used when the agents do not know *a priori* which actions are the best in order to achieve optimal performance. They make use of a variable called prediction error (PE) defined as the difference between the expected and the actual outcome [3].

Progress in the RL area was made, and evidences linking the PE signal with neural recordings and brain imaging data arose [4-7]. Now, the existence of a

correlation between PEs and dopamine (DA) is commonly accepted with positive PEs conveyed by DA burst and negative PEs conveyed by DA dips [6]. Also, plasticity of cortico-striatal synapses was found to be weighted by 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 [6, 8-10].

These ideas are combined in the basal ganglia Go/NoGo neurocomputational model [6, 11]. It is broadly agreed that basal ganglia (BG) is primarily implicated in selecting the best action to execute at a given time [12-14]. Whether by generating the actions or just selecting those being represented at the cortex, 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: the direct (Striatonigral or Go pathway) and the indirect (Striatopallidal or NoGo pathway). [6, 15]

If an action is followed by a dopamine burst (positive PE) the Go neurons in the striatum are strengthened, whereas the NoGo neurons are weakened. This leads to the disinhibition of the thalamus that, due to its excitatory connection with the motor cortex, will allow the occurrence of the appropriate movement. However, if an action is followed by a dopamine dip (negative PE) the opposite occurs and the output structures will no longer be disinhibited, which results in inhibition of the thalamus and constrainment of movements. [6, 15, 16].

There is still a third cortico-basal ganglia pathway, the hyperdirect pathway. This pathway has been shown to have particular relevance under conditions associated with response conflict [11] and preventing premature behavior [6, 16].

Although the relationship between PEs and dopamine is no longer questionable, the way these PEs are determined in the human brain is. RL presents different TD algorithms that can differ in how PEs are computed. For instance, in the SARSA (State-Action-Reward-State-Action) and in the Q-learning (QL) algorithms PEs are based on action values, while in the Actor-Critic (AC) model they made use of state values [3]. Since the central question in the present study concerns whether PEs in the human brain are driven by action values or by state values, the task used does not differentiate between the SARSA and the QL. We will just call these methods as the Q-learning framework.

Different studies have been supporting different models. At one hand, electrophysiological recordings (e.g.[17]) suggest a Q-learning framework. At the other hand, some studies point towards the AC approach suggesting that dopaminergic projections from VTA (Ventral Tegmental Area, the origin of the dopaminergic cell bodies), which targets the ventral striatum and other limbic areas, might be responsible for calculating the state values (Critic), while dopaminergic projections from SNc to dorsal striatum allows the policy learning (Actor) [18].

The AC model has also been supported by fMRI (functional Magnetic Resonance Imaging) 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 [19].

With all these divergent evidences it is of upmost importance to develop new studies that can give us a greater insight into the topic. One way to do it is by using Go/NoGo tasks – and since humans can be easily instructed, more complex scenarios can be considered. In this study we exposed the design of a new task, specially thought to address this question through instrumental conditioning (Go/NoGo) and classical conditioning (subliminal images). The task was tested in 35 healthy subjects and then the behavioral data was analysed.

Moreover, a deeper knowledge about this parallelism between RL algorithms and human decision-making processes can lead to great improvements in both areas:

i) 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;
ii) 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 [6].

## 2. Methods

### 2.1. Subjects

35 healthy adults (22 males and 13 females; age range 22-58 years; mean age $29.7 \pm 12.2$) participated in the

study, performing the task developed. 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.

### 2.2. Experimental design and Task

The task is divided into five phases: first there are two learning phases, with different trial types representing four different conditions; next a test phase; and finally the fourth and the fifth phases analyze the use of subliminal images shown during the learning phases.

#### 2.2.1. Learning Phases and Test Phase

Three of the four conditions used differ in their correct action (pressing/not pressing a button) and valence (win or lose) interaction: **go to win** (the subject has to press the button to win a reward); **go to avoid losing**, (the subject should press the button to avoid a punishment); and **nogo to avoid losing** (in order to avoid a punishment the subject needs to withhold from pressing the button). The fourth condition is named **neutral** because regardless the action performed the outcome is always zero. This condition will measure the bias to press the button.

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.

- In the **go to win** condition, pressing the button (Go trial) 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 (NoGo trial), the probability of receiving a neutral outcome is 80%, while receiving a reward has a probability of only 20%. In the **nogo to avoid losing** occurs exactly the opposite.
- In the **go to avoid losing** if the subject performs a Go trial he always receives a neutral outcome. However, if he does a NoGo trial with 80% probability he will get a punishment (-1); he can also have a neutral outcome with 20% probability.

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

The experimental paradigm used during both learning phases is presented in figure 1. In this phase subjects can choose between pressing or not the correspondent button, doing a Go or a NoGo trial, respectively. Then, in the 3rd phase participants are forced to choose one of the previous buttons and no feedback is given.

In the first learning phase four independent images are presented, each with a 25% probability: one representing a go to win condition (Distractor +, phase 1), other representing 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 in the 2nd phase, despite being associated with a neutral trial in this phase).

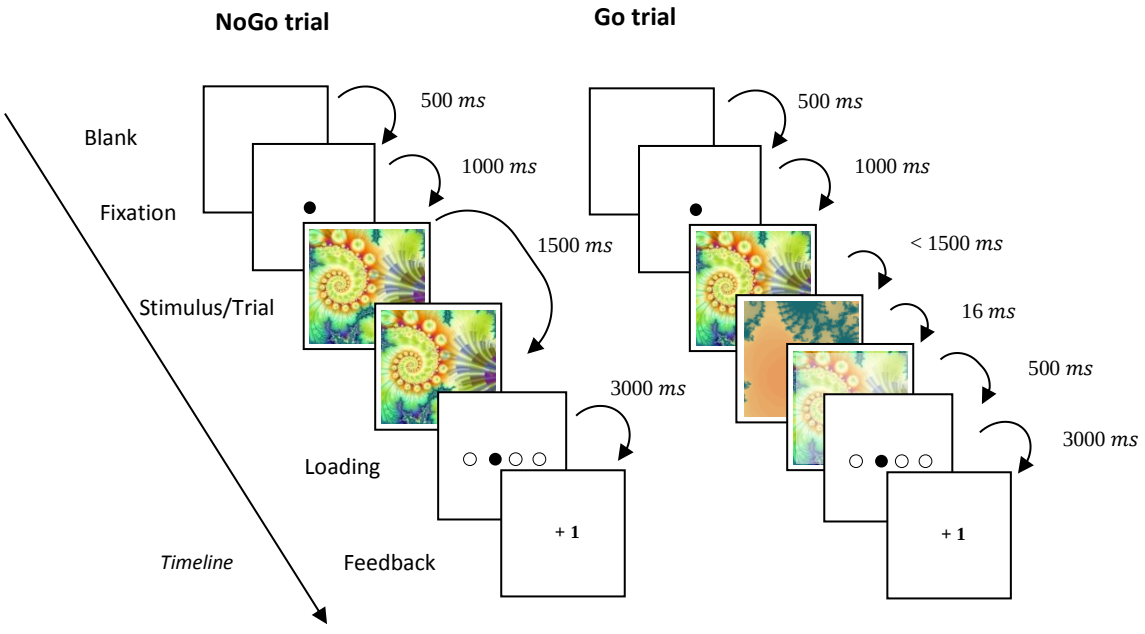


Figure 1: Experimental paradigm during the learning phases.

In the second learning phase 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 regardless 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 as in the previous phase.

This change in the go to avoid losing image from the 1st phase, where it acts as a neutral condition, to the 2nd phase, where it acts as a go to avoid losing condition, is the key aspect of the task. Due to the difference regarding the update of Q-values and preferences, there are now two possible expected outcomes for the go to avoid losing condition:

- If the QL 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);

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

  Here $\widehat{Q}^{\pi}(s_t, a_t)$ is the estimate of the Q-value $Q^{\pi}(s, a)$, the value of taking action $a$ in state s under policy π; α denotes the learning rate ($0 \leq \alpha \leq 1$); $\delta_t$ the prediction error and $r_{t+1}$ is the reward received in the transition of states.

- In the AC model the subject only computes preferences for Go actions, even though state-values $V^{\pi}(s)$, which represent the value of a state $s$ under a policy $\pi$, are updated in every trial regardless of the chosen action.
  So, if subjects follow the AC, they associate a higher preference value for button 2 (button of the 2nd phase). Then, it is expected that, when presented with the image and forced to make a choice, they will prefer to press button 2. The preference in the 1st phase associated with the go to avoid losing, like in the QL framework, was zero since it was acting as a neutral condition (with the outcome always being zero, $r_{t+1} = 0$). However, in the 2nd phase the state value of the image ($\hat{V}^{\pi}(s_t)$) acquires a negative value and so the prediction error in a Go trial will be positive, resulting in a positive preference for that action.

$$\hat{V}^{\pi}(s_t) \leftarrow \hat{V}^{\pi}(s_t) + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)] \quad (2)$$

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \eta\,\delta_t \quad (3)$$
$$where\ \delta_t = r_{t+1} - \hat{V}^{\pi}(s_t)$$

The go to avoid losing image is the most important condition in this study because is the only one capable of presenting this discrepancy between the two models occurs.

In any study is always necessary to minimize all the possible confounders (extraneous variables that may influence the results). So, during the task's design 3 major sources of confounders were studied:

i. **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.

ii. **Serial position effect** - The response tendency, either due to association or preference, might also depend on the serial position of each action. The subject could press button 2 not because he exhibits a preference for it but because it happened later on in the task (recency effect). The opposite effect could also happen (primacy effect). 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 (ensuring that it stays with a null value), before acting as a go to avoid losing condition (and therefore acquiring a value different from 0).
   Solution: Taking into account the subject's behavior in the Neutral image during the 3rd

phase, we can analyse if a subject is more biased to press a certain button.

iii. **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.

### 2.2.2. Subliminal Perception and Valence Phases

Every time the subject chooses to perform a go action (Go trial), a subliminal image is presented (a fractal image), immediately after pressing a button and before the trial image reappears (with a transparency layer) – see fig 1. Using subliminal images is another way to determine which model (Q-learning or AC) is being used by the brain during the learning process.

All stimulus images presented during the learning phases have a subliminal image associated. So, there are four subliminal images in the 1st phase and another four in the 2nd phase.

One essential point is making sure that those subliminal images are in fact subliminal. For that reason, the 4th phase is a *perceptual discrimination* task. Here sixteen different images were shown: 8 of them were the subliminal images displayed throughout the task, and the other 8 had never appeared before (control stimuli). The display order of the images was counterbalanced across subjects.

Subjects were asked to say if they had or not seen each image before. Then, the sensitivity index ($d'$) for each subject was calculated. This index is based on the difference between normalized rates of hits and false alarms [20]:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate) \quad (4)$$

So, d' values close to zero can be interpreted as a lack of conscious access during the main task.

In the fifth 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 mimics 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 behavior (through their choices during this evaluation phase), but fail to reach awareness in the direct d' test.

Similar to what was explained before regarding the test phase, analyzing the participants' behavior during this phase - their valence ratings – also give us insight about the calculation of PEs in the human brain.

The valence of an image can be seen as an indirect form of obtaining its state-value $V(S)_{subliminal}$ and again the subliminal image associated with to the **go to avoid losing** condition will be the key point.

$$\hat{V}^{\pi}(s_t)_{subliminal} \leftarrow \hat{V}^{\pi}(s_t)_{subliminal} + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)_{subliminal}] \quad (5)$$

$$r_{t+1} = r_{t+1_{go}} - 'what\ is\ expected' \quad (6)$$

The variable $'what\ is\ expected'$ can be either $\hat{Q}^{\pi}(s_t, a_{t,go})_{trial}$ or $\hat{V}^{\pi}(s_t)_{trial}$, according to which model is being used.

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

- 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** condition 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.

- In the AC case 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). Then, follows the valence associated with the Go to avoid losing image of the second phase, and finally, the valences 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:

$$\hat{V}^{\pi}(s_t)_{subliminal} \leftarrow \hat{V}^{\pi}(s_t)_{subliminal} + \alpha[r_{t+1} - \hat{V}^{\pi}(s_t)_{subliminal}] \quad (7)$$

$$r_{t+1} = r_{t+1_{go}} - \hat{V}^{\pi}(s_t)_{trial} \quad (8)$$

The reward for this condition, in a Go trial, is always zero, 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 reward is always zero the subject was expecting an aversive outcome $\hat{V}^{\pi}(s_t)_{trial} < 0$ that turned out to be neutral [21, 22].

# 3. Results

## 3.1. Learning Phases and Test Phase

Before turning our attention to 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, taking the neutral condition as reference.

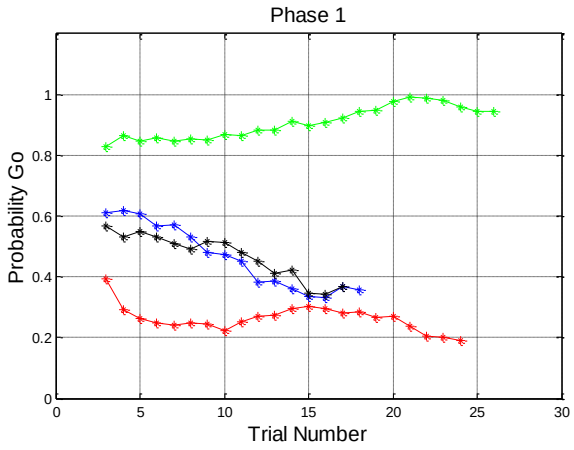


Figure 2: Time varying probabilities, across subjects of making a go response for each condition. Data was convolved with a central moving average filter with a length of 5 for the 1st phase. Data is only presented when there are at least 16 subjects for trial. In green is represented the positive distractor; in blue the go to avoid losing image; in black the neutral; and finally in red the negative distractor.

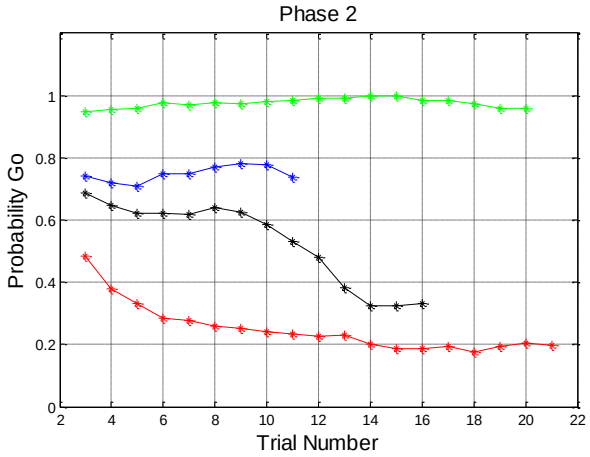


Figure 3: Time varying probabilities, across subjects of making a go response for each condition. Data was convolved with a central moving average filter with a length of 5 for the 2nd phase. Data is only presented when there are at least 16 subjects for trial. Same colors represent the same conditions, as in figure 2.

Even though the learning curves seem to confirm that subjects learnt correctly every condition, an ANOVA 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 in each phase with factor of condition.

In both phases, the ANOVA showed a main effect of condition: $F(3,102) = 3.556, p = 0.017$ in the 1st phase and $F(3,102) = 3.556, p < 0.001$ in the 2nd phase. However, we were mostly interested about the learning of the go to avoid losing condition. For that, a post hoc paired t-test between the **go to avoid losing** and the **neutral** images was done. In the first phase it did not revealed any significant difference ($p = 0.172$) – this was already expected since both images were acting as a neutral condition in this phase (fig. 4). However, in the second phase the post hoc paired t-test significantly showed that the go to avoid losing condition had a higher probability of Go trials which confirmed that correct learning had occurred ($p = 0.033$, fig. 5).

Regarding the test phase figure 6 shows us the percentage of button 2 choices for each trial image.

It can be seen that whenever a positive preference was associated with a certain button during a learning phase, in the Test Phase participants tended to press that same button (for the associated image).

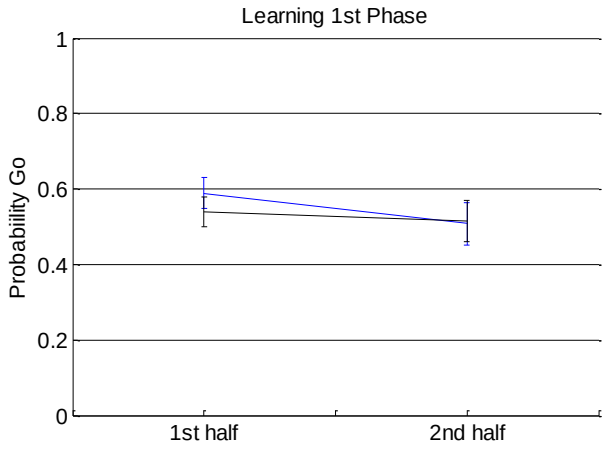


Figure 4: Probability of Go in the 1st and 2nd half for the go to avoid losing and neutral conditions, during the first learning phase. Error bars represent the standard error of the mean (S.E.M.). In blue is represented the go to avoid losing image and in black the neutral image.

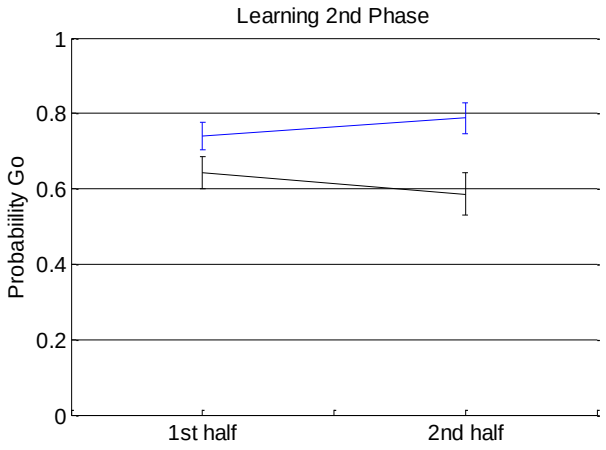


Figure 5: Probability of Go in the 1st and 2nd half for the go to avoid losing and neutral conditions, during the second learning phase. Error bars represent the standard error of the mean (S.E.M.). In blue is represented the go to avoid losing image and in black the neutral image.

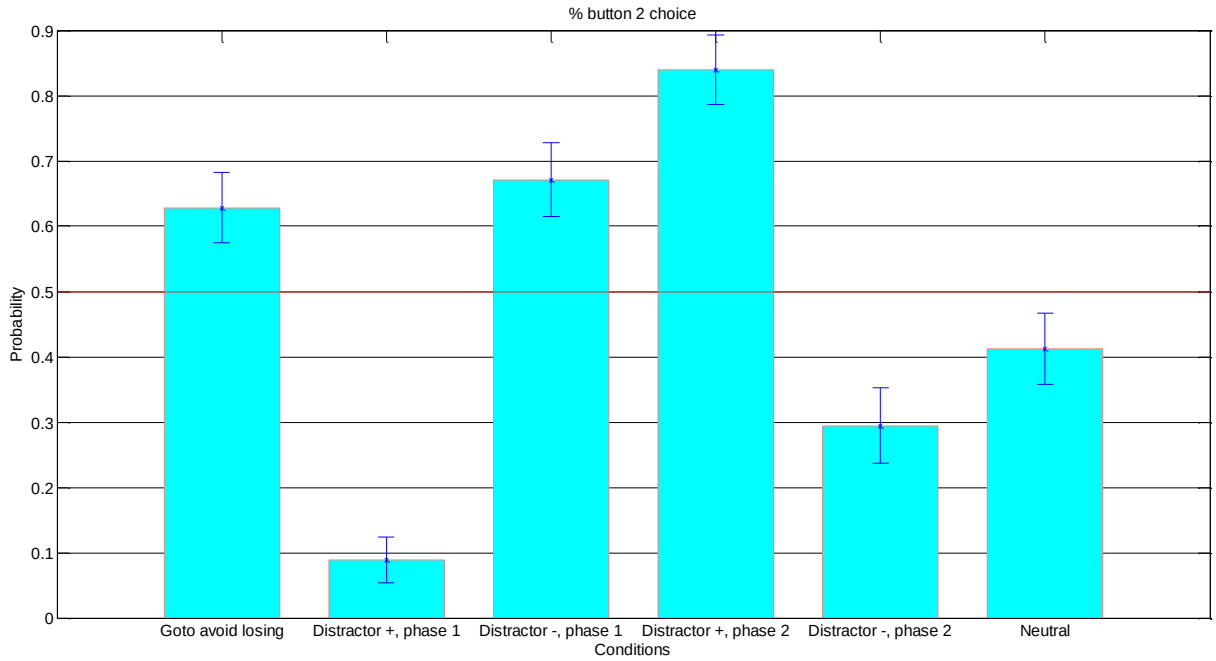


Figure 6: Percentage of button 2 choice in the 3rd phase for all conditions. Error bars represent the standard error of the mean (S.E.M.)

Subjects associated positive preference to press button 1 when the *Distractor +, phase 1* image appear in the 1st phase, so they were likely to press button 1 in the 3rd phase ($91{,}14 \pm 3{,}5\%$). The same reasoning can be made for the *Distractor +, phase 2* image, with participants showing a higher preference for button 2 ($84 \pm 5{,}30\%$).

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. Namely, subjects preferred to press button 2 ($67{,}14 \pm 5{,}58\%$) in response to the Distractor -, phase 1 image and button 1 ($70{,}57 \pm 5{,}42\%$) to Distractor -, phase 2 image during the 3rd phase.

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 preference – so, this bias needs to be taken into consideration. However, subjects have chosen almost equally the two buttons ($58{,}74 \pm 5{,}42\%$ the button 1 and $41{,}26 \pm 5{,}42\%$ the button 2) when the

stimulus image was the Neutral which indicates the inexistence of a serial position effect. The lower value might be explained by a cost to perform the pressing action.

Then, the tendency to press button 2 in the Go to avoid losing image ($62{,}86 \pm 5{,}35\%$) 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.

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 ($F(5{,}170) = 25.345, p < 0.001$). 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 ($p < 0.001$), 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

.

### 3.2. Subliminal Perception and Valence Phases

In the perceptual discrimination task only 5 of the 35 participants reported to have seen at least one of the displayed images. The *d'* values were calculated and using a one-tail paired t-test this measure was demonstrated not to be significantly different from zero, at group level ($p = 0.0506$).

The average of all given valences ratings, per image, is presented in figure 7.

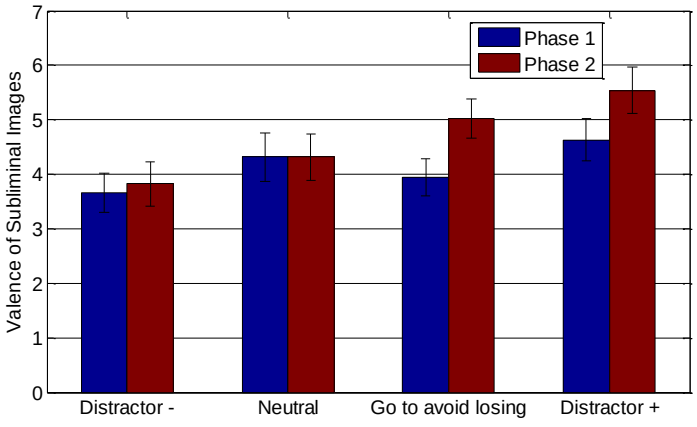


Figure 7: Subliminal images' valence. Error bars represent the standard error of the mean (S.E.M.).

Taking into consideration 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), 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 of values - 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. The result is depicted in figure 8.

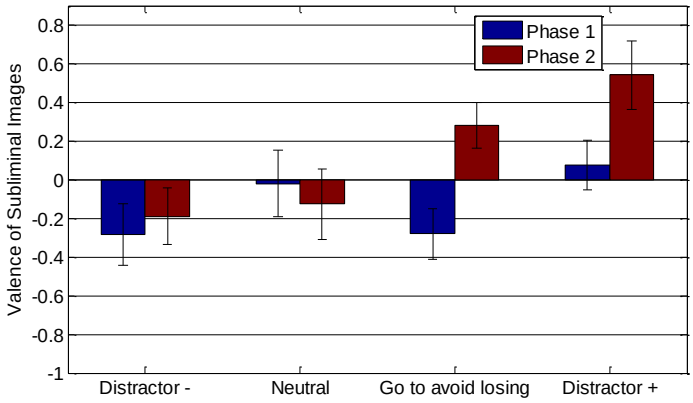


Figure 8: Subliminal images' valences normalized by the z-scores. Error bars represent the standard error of the mean (S.E.M.)

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

Once again, the difference of values between the Go to avoid losing's subliminal image of the 2nd phase and the Neutral's subliminal image of the same phase were statistically compared. A post hoc one-tailed paired t-test revealed that the Go to avoid losing assimilated a value statistical significantly higher than the Neutral image ($p = 0.0418$).

Additionally, it was also verified that the Go to avoid losing's subliminal image of the 2nd phase acquired a z-score valence significantly positive ($p = 0.0229$), while the Neutral's did not ($p = 0.4969$).

## 4. Conclusions

The aim of this study was to investigate whether prediction errors in the humans' brain are determined by state values (AC model) or by action-values (as in the 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 analysed.

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

was verified that participants correctly learnt the conditions during the learning phases and that subliminal images were in fact subliminal.

Behavioral data analyses from the two approaches were in agreement and strongly pointing towards the AC model.

On 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 condition. This supports the AC model ( $p(s,ago)_{button\,1}=0$), $p(s,ago)_{button\,2}>0$) , whereas a balanced behavior would point towards the Q-learning framework with$Q(s,ago)_{button\,1}=0; Q(s,ago)_{button\,2}=0$.

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 ($\mathrm{V(S)_{subliminal}}$). 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 AC model, since throughout time, for the go to avoid losing's subliminal image $\hat{V}^{\pi}(s_t)_{subliminal} \rightarrow \hat{V}^{\pi}(s_t)_{subliminal} > 0$ , while for the neutral's it remains zero ($\hat{V}^{\pi}(s_t)_{subliminal}=0$). If the Q-learning had been the model employed, both state-values would have been kept neutral.

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) [19], while others may use them to learn an action-selection policy (being the Actor). fMRI studies also support this theory (e.g. [18]) and so, they are in line with our findings. Furthermore, the AC model is also biologically congruent with the basal ganglia Go/NoGo (BG-GNG) model [6, 11] which strongly supports our results.

## Acknowledgements

I would like to thank my supervisor, my amazing group at the lab, my friends all over the world and specially my family for all the support.

## References

1. Domjan, M., *Principles of Learning and Behavior Active Learning*. 6 ed. 2010: Cengage Learning.
2. Thorndike, E.L., *Animal inteligence: Experimental studies.* New York: Macmillan, 1911.
3. Sutton, R.S. and A.G. Barto, *Reinforcement Learning: an Introduction*. 1998: The MIT Press.
4. 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.
5. 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.
6. Maia, T.V. and M.J. Frank, *From reinforcement learning models to psychiatric and neurological disorders.* Nat Neurosci, 2011. 14(2): p. 154-62.
7. Shah, A., *Psychological and Neuroscientific Connections with Reinforcement Learning*, in *Reinforcement Learning: State of the Art*, M. Wiering and M.v. Otterlo, Editors. 2012.
8. Hebb, D.O., *The organization of behavior*. . New York: Wiley, 1949.
9. Reynolds, J.N., B.I. Hyland, and J.R. Wickens, *A cellular mechanism of reward-related learning.* Nature, 413.6851 (2001): 67-70.
10. 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.
11. 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.
12. Wickens, J.R., et al., *Dopaminergic mechanisms in actions and habits.* J Neurosci, 2007. 27(31): p. 8181-3.
13. Albin, R., A. Youg, and J. Penney, *The functional anatomy of basal ganglia disorders.* Trends in Neuroscience, 1989. 366-375.
14. Mink, J., *The basal ganglia focused selection and inhibition of competing motor programs.* Progress in Neurobiology, (1996): 381-425.
15. Gerfen, C.R., *Molecular effects of dopamine on striatal-projection pathways.* Trends in Neuroscience, 23 (2000): S64-S70.
16. Nambu, A., H. Tokuno, and M. Takada, *Functional significance of the cortico–subthalamo–pallidal 'hyperdirect'pathway.* Neuroscience research 43.2 (2002): 111-117.
17. 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.
18. 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.
19. 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.
20. Vermeiren, A. and A. Cleeremans, *The Validity of d' Measures.* Plos one, 2012.
21. Seymour, B., et al., *Differential encoding of losses and gains in the human striatum.* J Neurosci, 2007. 27(18): p. 4826-31.
22. Tanimoto, H., M. Heisenberg, and B. Gerber, *Even timing turns punishment to rewards.* Nature, 2004. 430: p. 983.