Mechanistic characterization of reinforcement learning in healthy humans using computational models

Ângelo Rodrigo Neto Dias
Under supervision of Tiago Vaz Maia

July 14, 2014

Abstract

Reinforcement learning has provided a normative framework to analyse decision-making. A wealth of research has linked reinforcement learning to neural substrates, assigning them a particular computational role. Particularly, responses of dopamine neurons can be identified with the prediction errors computed in the temporal-difference learning algorithms. Machine learning literature has proposed different versions of calculating the error signal, associated with different temporal-difference algorithms. Particularly, they can be determined by the value of actions (Q-learning model) or by the value of states (Actor-Critic model). Neuroscientific findings have supported both models, and thus, there is still no commonly accepted mechanism.

The aim of this thesis was to investigate and identify which of these two reinforcement learning models best describes the choices made by healthy humans when performing a modified probabilistic Go/NoGo task. This paradigm has the special feature of orthogonalizing action and valence and thus it enhances some mechanistic differences between the Q-learning and the Actor-critic models.

For this purpose, we employed several statistical methods. Firstly, using a model fitting approach we tried to identify which of the aforementioned models best suited data. Secondly, we performed a Principal Component Analysis in order to find associations among conditions which could also provide evidence towards one of the models.

Both approaches provided evidence towards the Q-learning framework which indicated that the prediction errors are determined by the value of actions. This result was in line with electrophysiological findings in animals.

Keywords: Actor-Critic, Q-Learning, Prediction Errors, Dopamine, Go/NoGo task

1 Introduction

A fundamental question in behavioral neuroscience concerns the decision-making process used by animals and humans for selecting actions in the face of reward and punishment. This process has been extensively investigated through the paradigms of classical and instrumental conditioning.

In classical conditioning subjects learn the association between events [1]. In contrast, instrumental learning involves learning the association between actions and events [2, 3]. According to the Thorndike's law, in the presence of a reward the connection is strengthened and in the presence of a punishment the connection is weakened [2].

This trial and error procedure is also found in temporal difference (TD) reinforcement learning algorithms which are commonly employed in artificial systems, such as robots, to make them capable of learning to select actions [4, 5]. Therefore, they have gained popularity in behavioral neuroscience to explain conditioning behavior. These models learn how to make a decision by predicting the value of taking an action from a recognized state. The subjects can thus choose the action which maximizes that value. The value is then updated through the prediction error (PE) defined as the difference between the expected and the observed value.

In the last decade, several observations showed a good parallel between neurobiological processes in the brain and the computational steps of Reinforcement learning algorithms. The most notable finding was the relationship between dopamine (DA) and PEs. It was proved that phasic responses of the midbrain dopamine code reinforcement PEs. Namely, positive PEs are conveyed by DA bursts and negative PEs are conveyed by DA dips [6–8]. Another important evidence came from the studies done by Wickens and colleagues who found that the plasticity of corticostriatal synapses is weighted by dopamine input from mid-

brain dopamine neurons [9–12]. Additionally, some findings suggest that striatum plays a major role in linking the value to action selection [13–15].

These observations are unified in the basal ganglia Go/NoGo neurocomputational model [16]. Anatomically, the Basal ganglia (BG) exhibits two main pathways: the direct and indirect pathway. The direct pathway facilitates movement whereas the indirect pathway suppresses movement. According to the basal ganglia Go/NoGo model, when an action is followed by a positive PE, a phasic DA burst occurs which will strengthen the direct pathway and weaken the indirect pathway. When a negative PE occurs, the opposite happens [16, 17]. The tonic DA increases and decreases the excitability of the direct and indirect pathways, respectively. This way, high levels of tonic dopamine increases the tendency to respond (Go bias) and reduces the tendency to not respond (NoGo bias), and vice-versa [16, 18].

Machine learning literature has proposed different versions of calculating PEs, associated with different TD algorithms (e.g. [4]). In the Actor-critic algorithms [19], the PE ignores actions altogether, and thus, it is determined by the value of the situation. In the other two classes of algorithms, Q-learning (QL) [20] and SARSA (state-action-reward-state-action) [21], PEs are determined by the value of the action (Q-value). In the QL approach, the PEs associated to a decision are determined by the Q-value of the better option rather than the one actually chosen. On the other hand, in SARSA algorithms the PEs use the Q-value of the chosen option.

Neuroanatomical findings support the Actor-critic (AC) model [19], whereas electrophysiological evidences are in line with the other two classes of algorithms. Recent evidence from primate study seems to support SARSA [22]. On the other hand, evidence from a rodent study favours QL [23]. Therefore, resolving this discrepancy necessitates further experiments and computational investigation.

Although animals can display complex decision-making behavior, we are interested in comprehend human-decision making and its relationship with the Reinforcement learning (RL) framework. Furthermore, the possibility of instructing subjects verbally allows for much more complex paradigms in human experiments.

This way, we used a task which highlights some differences between the AC and QL framework, in order to determine whether PEs, in humans, are governed by the value of the situations (AC) or by the value of the actions (QL and SARSA). This work is not concerned about the distinction between QL and SARSA models, and thus, we will just consider the class formed by the QL and SARSA algorithms, instead of considering them separately. Since both approaches involve Q-values, this class will, henceforth, be denominated as Q-learning.

Recent studies have used tasks that fully orthogonalize action and valence in a 2 (reward/punishment) $\times$ 2 (Go/NoGo) design [24, 25]. This type of task highlights a gap between the Q-learning and the basal ganglia Go/NoGo model. Typically, in this kind of task, there are four different conditions: respond to gain a reward (go to win); respond to avoid punishment (go to avoid losing); do not respond to gain a reward (nogo to win); do not respond to avoid punishment (nogo to avoid losing). Assuming a QL approach, the nogo to win condition exhibits positive PEs and the go to avoid losing exhibits negative PEs, and thus they are linked to dopamine bursts and dips, respectively. According to the basal ganglia Go/NoGo model, the nogo to win condition would strengthen the direct pathway, and thus, promoting the go action, instead of suppressing it. Conversely, the go to avoid losing condition would weaken the direct pathway and strengthen the indirect pathway, which would promote the nogo action. This incongruence reveals a gap between the QL algorithm and the neurobiological explanation, which can be overcame by the AC model.

The AC model is in line with the basal ganglia model. The nogo to win produces negative PEs whereas the go to avoid losing produces positive PEs, and thus, they strengthen the indirect and weaken the direct pathway, which promotes the nogo action.

A recent study [25], which used a model whose structure is identical to the QL, showed that subjects perform worse in the nogo to win and go to avoid losing conditions. They explained these findings by assuming the existence of a Pavlovian effect, i.e. the subjects tend to respond in conditions whose outcome is positive (nogo to win and go to win) and tend to not respond when the outcome is negative (go to avoid losing and nogo to avoid losing). We suggest that this action by valence interaction might be due to the fact that subjects execute the AC model instead of the QL. The AC model makes the learning process slower in the nogo to win and go to avoid losing conditions, which could explain these findings.

2 Material and Methods

2.1 Subjects

24 adults participated in this experiment (10 females and 14 males; age range 19-50 years; mean=25.38, SD=7.82 years). All the participants performed the task voluntarily.

2.2 Experimental design and Task

The task has five different type of conditions represented by five fractal images: press the key to gain a reward (go to win); do not press the key to gain a reward (nogo to win); press the key to avoid a pun-

ishment (go to avoid losing); do not press the key to avoid a punishment (nogo to avoid losing). In the fifth condition the subject always receives a neutral outcome regardless of the action performed (neutral). The neutral condition measures the Go bias, which is important to determine whether subjects have actually learned, by comparing this baseline behavior to the behavior in the other conditions.

The outcomes are presented stochastically and the probability distributions for each stimulus are described in figure 1.

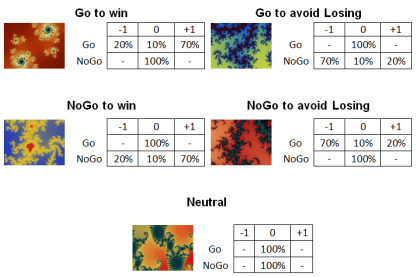

Figure 1: Probability distribution of the outcomes for the go to win, nogo to win, go to avoid losing and nogo to avoid losing conditions. The possible outcomes are $-1, 0, 1$, which corresponds to reward, neutral and punishment, respectively.

The task is divided into 3 blocks, each with 10 trials of each condition. All trials share the same structure. At the begin of every trial, a fixation point is displayed on the screen for 1000 ms. After this, a fractal image appears on the screen for 1500 ms. During this period, the participants had to choose whether to press or not the key. After the offset of the image, there is a variable interval (3000 ± 1000 ms) during which a progress indicator is displayed in order to give the participant a sense of time until he/she waits for the feedback. Furthermore, it tries to avoid the participant from keep on pressing the key. Participants are presented with the feedback after this interval. The feedback remains on the screen for 1000 ms: a green +1 indicates a win of 1 point (reward), a red -1 indicates a loss of 1 point (punishment) and a black 0 indicates no win or loss (neutral). The feedback is followed by a blank screen which takes 3000 ± 2000 ms. After this blank period, the next trial starts.

Images and conditions were counterbalanced across subjects in order to avoid interactions between the images and the conditions. The order of appearance of the images was randomized within blocks. The task was run in a HP Pavilion dm4 Notebook PC with a 14-inch screen.

3 Behavioral analysis

To characterize the group performance in the task, we have done two different approaches. The first approach did not consider the Go bias measured by the neutral condition and the second did.

In both analysis we have done a repeated measures three-way ANOVA for the number of correct responses, with factors of block (3 levels), valence (win/avoid losing: 2 levels) and action (Go/NoGo: 2 levels), and a one-way ANOVA with factors of block (3 levels). The number of correct responses corresponds to the number of times the subjects pressed in the go conditions and to the number of times they did not press in the nogo conditions. If considering the Go bias, the number of correct responses are subtracted by the Go and NoGo bias in the go and nogo conditions, respectively. This measure was called adjusted correct responses. Both measures were collapsed into bins of 10 trials per condition. This statistical analysis was performed in SPSS® version 16.

4 RL models

4.1 QL models

We built 4 parametrized QL models to fit to the behavior of the subjects. All models assigned a probability to each action a_t on trial t according to the action weight $W(s_t, a_t)$. This was based on the softmax method [4]:

$$P(a_t|s_t) = \frac{e^{\beta W(s_t,a_t)}}{\sum_{b \in A(s)} e^{\beta W(s_t,b)}} \qquad (1)$$

where parameter $\beta \geq 0$ assigned some degree of randomness in the subjects' choices. The states s_t corresponded to the five conditions presented in the task and a_t corresponded to the 2 possible actions: respond (go) or not respond (nogo).

The models further differed in terms of how action weight was constructed. For the standard QL model, $W(s_t, a_t) = Q(s_t, a_t)$, which was updated recursively according to the equation 2.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t - Q(a_t, s_t)] \qquad (2)$$

The parameter $0 \leq \alpha \leq 1$ was the learning rate and it dictated how fast the agent learned. The feedbacks entered the equation through $r_t \in \{-1, 1, 0\}$.

The standard model assumed that the initial Q-value was zero, and thus both actions (go and nogo) were equally probable. However, subjects might exhibit a natural bias towards the go or nogo action. The model QL+Q_0 captured this initial bias by allowing the Q-value on the first trial to vary freely between -1 and 1, while for all other models this was set to zero.

This tendency to perform an action can also be taken into account by a bias parameter. Unlike Q_0, which is gradually erased as new feedbacks are integrated, the bias parameter remains constant across

the experiment (equation 3).

$$W(s_t, go) = Q(s_t, go) + b \tag{3}$$

In order to capture the action by valence interaction, a Pavlovian parameter was included in the action weight (equation 8).

$$W(s_t, go) = Q(s_t, go) + \pi V(s_t) \tag{4}$$

The parameter π was constrained to always be positive. Thus, for conditions in which the outcomes were mostly negative ($V(s_t) < 0$), the Pavlovian factor decreased the tendency to go, while it promoted the tendency to go in conditions where the outcomes are mostly positive ($V(s_t) > 0$).

The state-value function was evaluated according to the equation 5.

$$V(s_t) \leftarrow V(s_t) + \alpha[r_t - V(s_t)] \tag{5}$$

4.2 AC models

For the AC framework, we built 4 parameterized learning models. In contrast to the QL models, the AC framework uses the state-values to compute the PEs and it is divided into two structures: the critic and the actor. The former selects actions and the latter criticizes the policy currently followed by the actor using the PEs. In this model, only the go action weight was considered (equation 6).

$$\begin{array}{rcl} P(go|s_t) = \pi(s_t, go) & = & \frac{e^{\beta W(go,a_t)}}{e^{\beta W(go,a_t)}+1} \\ P(nogo|s_t) = \pi(s_t, nogo) & = & \frac{1}{e^{\beta W(go,a_t)}+1} \end{array} \tag{6}$$

For the standard AC model, $W(s_t, go) = p(s_t, go)$, which was updated recursively according to the equation 7.

$$p(s_t, go) \leftarrow p(s_t, go) + \eta[r_t - V(s_t)] \tag{7}$$

$0 \leq \eta \leq 1$ was the actor's learning rate. The critic updated the state-value function according to the equation 5.

The standard AC model did not take into account that subjects might exhibit a go bias. Consequently, we tested the model AC+p_0 which captured this bias by allowing the preferences on the first trial to vary freely. Unlike Q_0 in the QL framework,p_0 remained static across time.

For the model including the Pavlovian factor (AC+*pav*), the action weight was modified in the same way as in the QL+*pav* model,

$$W(s_t, go) = p(s_t, go) + \pi V(s_t) \tag{8}$$

5 Model fitting procedure

The previous models were fitted to subjects' behavioral data. This was achieved by the Maximum likelihood estimation (MLE) method. Given the models described above (M) and the sequence of actions made by subject s along T trials (i.e. $D_s = a_{s_1}, ..., a_{s_T}$), the likelihood of the entire dataset D_s is just the product of their probabilities from equations 1 or 6,

$$P(D_s|\theta_s, M) = \prod_{i=1}^{T} P(a_{s_i}|\bigcap_{j=1}^{i-1} a_{s_j}, \theta_s, M) \tag{9}$$

We estimated the free parameters θ_s of model M by seeking the set of parameters which maximized the Log-Likelihood (LLH)[1] function. This optimization was performed in MATLAB, through the fmincon routine, using a set of different starting points widely dispersed over the search domain.

6 Model Comparison

The individual model goodness was measured using the Bayesian Information Criterion (BIC) [26], which is an approximation of the log-model evidence (equation 10).

$$\log P(D_s|M) \sim \log P(D_s|M, \theta) - \frac{n}{2} \log m \tag{10}$$

The models were compared at the group level using both classical and Bayesian approaches [27]. Since the Bayesian comparison is a more recent approach than the classical analysis, and thus, it has been less employed in model comparison, we decided to supplement it with the classical approach.

In the classical setting one uses the log-model evidence across subjects, testing the null hypothesis that one model is no better than the other. This is accomplished by performing a simple one-sample t-test on the log-model evidences. The t-test assumes that the log-model evidences are normally distributed, thus a test of normality must be performed. Here, we adopted the Kolmogorov-Smirnoff test to test for this parametric assumption. Whenever this test rejects the null hypothesis of a normally distributed data, we use a Wilcoxon signed rank test which does not make any distributional assumptions [27]. This classical random effects approach can generate incorrect results when the inter-subject variability is high due to outliers. A more robust method to outliers was recently presented which consists in a Bayesian framework [27].

[1] Because of the monotonic nature of the logarithm, one can maximize the likelihood or the log-likelihood; the latter, however, is numerically more convenient to deal with.

This approach is based on a hierarchical Bayesian model, where each subject model is sampled from a multinomial distribution and then individual data is generated under that subject-specific model.

Thereby, the multinomial distribution parameters (r) are the probability of each model in the population. The model probabilities follows in turn a Dirichlet distribution with parameter α. In order to compare models, we are interested in the density from which models are sampled to generate subject-specific data. In other words, we seek the probability of the multinomial parameters given the data, i.e. $P(r|D_s)$. This can be achieved by inverting the hierarchical Bayesian model using an variational Bayes approach [27]. This method was implemented using the *spm_BMS* function from the SPM8 toolbox in MATLAB. The density over r can be used to quantify our belief that a particular model $k1$ is more likely than other model $k2$, given the observed data, denominated as exceedance probability (ϕ) (equation 11)

$$\phi_{k1} = P(r_{k_1} > r_{k_2}|y) \tag{11}$$

Since $r_{k_1} + r_{k_2} = 1$, equation 11 can be re-written as $\phi_{k1} = P(r_{k_1} > 0.5|y)$.

The exceedance probability is a statement of belief about the posterior probability, not the posterior itself. So, for example, when the exceedance probability is 98%, it means that we can be 98% confident that the favoured model has a greater posterior probability than any other model tested.

7 Results

7.1 Behavioral results

The optimal choice in the go to win and go to avoid losing conditions is to go. Conversely, the optimal choice is not to emit an action in the nogo to win and nogo to avoid losing conditions. Figure 2 shows the average choice probabilities for all subjects. The learning curves for each of the five conditions suggest that subjects have learned at most three conditions: go to win, go to avoid losing and nogo to avoid losing.

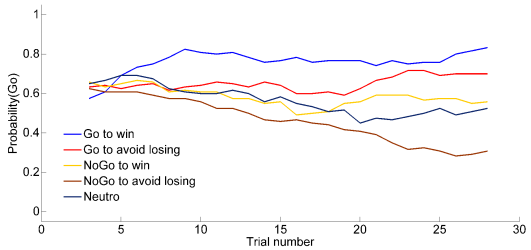

Figure 2: Average time varying probability, across subjects, of making the go action for the five conditions convolved with a central moving average with length 5.

A three-way ANOVA on the number of correct responses with factors of block (3 time bins of 10 trials each), action (go/nogo) and valence (win/avoid losing) revealed a main effect of block ($F(2,46) = 10.742, p < 0.001$). However, the presence of a block by valence interaction ($F(2,46) = 4.643, p = 0.015$) did not allow us to determine whether all conditions were effectively learned. A one-way ANOVA with factors of block for each condition showed a main effect in the go to win ($F(2,46) = 4.031, p = 0.024$) and nogo to avoid losing ($F(2,46) = 14.506, p < 0.001$) conditions. However, it did not exhibit a main effect in the nogo to win and go to avoid losing conditions. These findings are in line with the Pavlovian effect previously described: subjects tend to respond in conditions whose outcome is positive (nogo to win) and to not respond when the outcome is negative (go to avoid losing). Indeed, the three-way ANOVA test exhibited an action by valence interaction ($F(1,23) = 11.581, p = 0.002$). A post hoc paired t-test revealed a greater number of correct responses in the second block in go to win than in the go to avoid losing condition ($t(23) = 2.057, p = 0.051$)[2]. The nogo to avoid losing condition showed a significant greater number of correct responses than the nogo to win condition in the third block ($t(23) = -3.784, p = 0.001$). Thus, the behavioral data indicates that subjects were better at learning to go in the win condition (compared to go in the avoid losing), and were better at learning to withhold a response in the avoid losing condition (compared to not respond in the win condition). Nevertheless, subjects learned correctly the go to avoid losing conditions, although reached the asymptote quicker in the go to win.

Using the number of adjusted correct responses, the go to win ($F(2,46) = 5.5, p = 0.007$) and go to avoid losing ($F(2,46) = 6.154, p = 0.004$) conditions exhibited a main effect of block, whereas the nogo conditions did not. Although, these results indicated that subjects did not learn the nogo to avoid losing condition, this condition exhibited a positive trend (post-hoc paired t-test between the first and third blocks, $t(23) = -1.776, p = 0.089$) whereas the nogo to win showed a negative trend. The rest of the results were similar to the previous analysis, and thus, the conclusions were the same.

7.2 Comparing QL to AC

In order to find the model which best fits to the behavioural data, we compared several QL and AC models. Firstly, we fitted the standard QL model, which was purely instrumental with two parameters: the learning rate (α) and the inverse of temperature (β).

The subjects exhibited some initial tendency to respond which was impossible to capture using a stan-

[2]Since this result showed a p-value very near the threshold, it is worth to also take it into consideration.

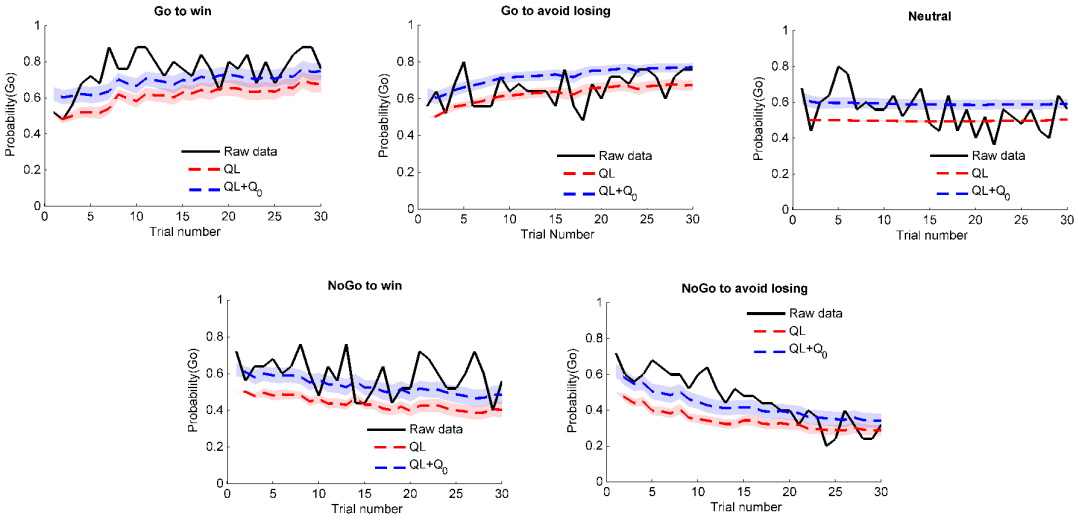

Figure 3: Learning time courses for all five conditions. The black lines depict the time varying probabilities, across subjects, of making a go response. The red lines represent the same average time varying probabilities, across subjects, but sampled from the standard QL model and the blue lines were sampled from the QL+Q_0 model.

dard Q-learning model (red lines in figure 3, the initial probability of the go action was above 0.5 in most of the conditions, namely go to avoid losing, neutral, nogo to win and nogo to avoid losing). Consequently, we tested for two alternative models to account for this effect: firstly, we included an initial Q-value (QL+Q_0), which gradually vanishes as new outcomes are integrated; secondly, we included a bias parameter which remains static across time. The Bayesian analysis favoured the latter, but the classical approach did not (Wilcoxon signed rank test: $p = 0.2769$, Bayesian: $\phi = 0.8160$). Indeed the model's simulated behavior were practically the same in both models.

However, the QL+Q_0 model slightly failed to capture the tendency to respond in the win conditions and the tendency to not respond in the avoid losing conditions. Particularly, this failure was more pronounced in the go to avoid losing condition (blue lines in figure 3). Thus, we tested the QL+pav model which accounted for the action by valence interaction. Although the classical approach was not able to determine which was the best model ($t(23) = -1.4433, p = 0.1624$), the Bayesian model comparison (BMS) method supported strongly the model without the Pavlovian factor with an exceedance probability of $\phi = 0.9964$.

Regarding the AC framework, we firstly fitted a standard AC model which consisted of three parameters: the critic's learning rate (α), the actor's learning rate (η) and the inverse of temperature (β).

Similarly to the QL framework, the standard AC model failed to capture the initial tendency to perform the go action. Therefore, we tested an alternative model which included an initial preference (AC+p_0), ans thus, it was able to account for this effect. This model was clearly better than the standard one (Wilcoxon signed rank test: $p = 0.0125$, Bayesian: $\phi = 0.9957$).

Although, the AC+p_0 model seemed to capture the action by valence interaction, we still fitted the AC+pav model to the behavioral data and compared it to the AC+p_0 model, so that we could be sure that the Pavlovian factor was useless. Indeed, the Pavlovian factor was not required to explain data (Wilcoxon signed rank test: $p < 0.001$, Bayesian:$\phi = 1$).

Thus, our computational analysis suggested that the QL+Q_0 and AC+p_0 models were the best candidates to explain the subjects' behavioral data.

The AC+p_0 model was then compared to the QL+Q_0. This comparison showed that QL model had a higher probability of occurring in the population with an exceedance probability of $\phi = 0.8780$.

Although our findings suggested that the subjects' behavioral data is in line with the QL model, the AC model might naturally capture better the Pavlovian effect. In the AC framework, subjects must learn the state-value before making a decision. In the nogo to win and go to avoid losing conditions, in order to learn it, subjects must perform the nogo action, which does not update the preferences and thus it slows the learning process. The Pavlovian effect is characterized by a slower learning process in these two conditions. Therefore, we hypothesized that it could be better captured by the AC model.

In order to verify this hypothesis, we performed a two-way ANOVA for the probability of correct responses sampled from the QL+Q_0 and the AC+p_0 models, with factors of action (2 levels: go and nogo) and valence (2 levels: win and avoid losing). This analysis showed that the AC model ($F(1, 23) = 19.595, p < 0.001$) showed an interaction action-

Table 1: Prediction errors underlying each condition in the Q-learning and Actor-critic models.

	Go to win	NoGo to win	Go to avoid losing	NoGo to avoid losing
QL	$\delta \geq 0$	$\delta \geq 0$	$\delta \leq 0$	$\delta \leq 0$
AC	$\delta \geq 0$	$\delta \leq 0$	$\delta \geq 0$	$\delta \leq 0$

valence more similar to the raw data than the QL ($F(1, 23) = 4.590, p = 0.043$).

7.3 Principal Component Analysis of behavioral data

The Actor-critic and the Q-learning models differ in the sign of the PEs in the nogo to win and go to avoid losing conditions. For example, consider the nogo to win condition (respond: positive outcome, not respond: negative outcome): in a QL framework, as subjects learned the action-values during the task, PEs become positive ($0 \leq Q(s, nogo) \leq 1, \delta = 1 - Q(s, nogo) > 0$) when they withhold their responses. In an AC framework, the PEs are determined by the state-value function and the preferences are only updated when the subjects respond. Therefore, as subjects learned the state-values during the task, PEs became negative ($V(s) > 0, \delta = 0 - V(s) < 0$) every time they respond. The dissimilarities in prediction errors according to each model are summarized in table 1.

As previously described, positive PEs are conveyed by dopamine bursts in the striatum, whereas negative PEs are conveyed by dopamine dips. Dopamine affects the subjects' learning performance [28–30], namely augmented DA enhances learning from positive PEs, but would impair learning from negative PEs. Conversely, reduced DA would enhance learning from negative PEs, but would impair learning from positive PEs.

Therefore, we hypothesized that conditions might be associated with each other in terms of performance (correct or incorrect learning). In a Q-learning framework, the win conditions, which are determined by positive PEs (DA bursts), would be positively correlated. Similarly, the avoid losing conditions, which are determined by negative PEs (DA dips), would exhibit a positive correlation. Additionally, the win and the avoid losing conditions would be negatively correlated.

In an AC framework, the go conditions, which are governed by positive PEs (DA bursts), would show a positive association. The performance in the nogo conditions, which are controlled by negative PEs (DA dips) would also exhibit a positive association. Additionally, the go and nogo conditions would be negatively correlated. Thereby, in order to investigate which model, QL or AC, is in line with the subjects' behavior, we tried to find associations among conditions in the behavioral raw data. This was achieved by performing a principal component analysis (PCA) on the subjects' performance in each condition. The subjects' performance was mathematically formulated as the fraction of correct responses in the third block.

Firstly, to verify the presence of correlations among conditions, we computed the Pearson's correlation coefficients. This preliminary analysis showed that only the go to avoid losing and the nogo to avoid losing conditions exhibited a positive strong and significant association ($r = 0.664, p = 0.001$)[3].

The principal component analysis was performed on the the same dataset. The principal components are depicted in the figure 4. Before proceeding with the analyses of the results, we determined the more meaningful components.

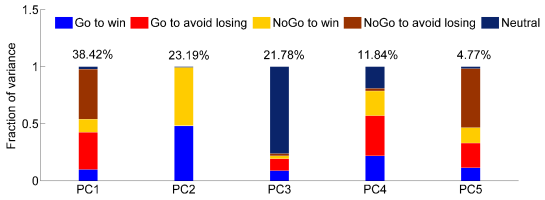

Figure 4: Results of the PCA analysis on the fraction of correct responses per subject and condition. Each set of stacked bars represents the fraction of the variance of a component explained by conditions. The percentage on the top of each stacked bars corresponds to the contribution of each component to explain the total variance of the original data.

In order to find the meaningful components we combined the K1 method and the Cattel's scree test. According to the former, all the components whose eigenvalues were above 1 should be retained whereas the others should be discarded. The first three components exhibited eigenvalues above 1 and, thus, they were considered significant. Additionally, the scree plot showed a significant last break at the third principal component, and, thus, according to the Cattel's scree test, only the first three components should be retained. This means that both tests agreed in retaining components 1,2 and 3. Our analysis, hence, only focused on these three principal components.

[3]The strength of the Pearson correlation coefficient was categorized according to the criteria introduced in [31] as strong ($r = \pm.10$ to $\pm.29$), moderate ($r = \pm.30$ to $\pm.49$) and weak ($r = \pm.50$ to ±1.0).

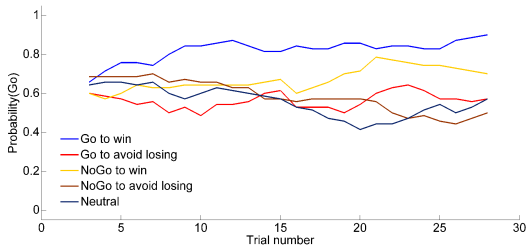

Figure 5: sub-population 1

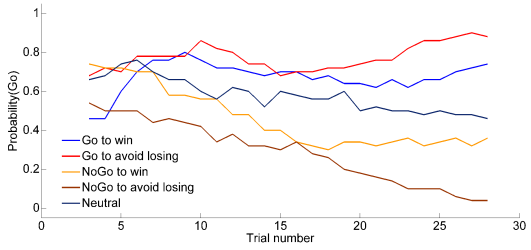

Figure 6: sub-population 2

Figure 7: Average time varying probability, across subjects, of making the go action for the five conditions convolved with a central moving average filter with length 5.

The results showed that the principal components grouped together the conditions which share the same sign of the PEs in the Q-learning framework: the avoid losing conditions are connected to negative PEs (first component), the win conditions are linked to positive PEs (second component), and the neutral are connected to zero PEs (third component). Additionally, the vectors associated to the first two principal components showed a positive association in the win and avoid losing conditions.

According to what was explained above about the relationship between dopamine and prediction errors, it would be plausible to expect that the avoid losing and win conditions shared the same driving force (same component), i.e. level of dopamine in the striatum. As previously explained, there were evidences which suggested that negative prediction errors might be conveyed by the dip duration instead of its magnitude. This could be the reason why win and avoid losing conditions showed a different driving force, and, thus, belonged to orthogonal components. The neutral condition measures the Go bias. There were evidences that this bias depended on tonic dopamine instead on the phasic dopamine. It is still unknown if there is a relationship between phasic and tonic dopamine. However, these results suggest that they are uncorrelated. The second component showed an association between the conditions, but the Pearson's correlation analysis exhibited a weak and non-significant correlation ($r = .159, p = .459$). Both analyses assumed that all the subjects performed solely according to one policy, but subjects could exhibit a more Pavlovian or more instrumental policy. Thereby, we should have accounted for this difference in policy, in order to capture correctly the associations among conditions.

To identify the presence of these two sub-groups in the overall population, we carried out a Gaussian mixture model on the 3-dimensional space spanned by the three components. This analysis was performed in MATLAB version 2012b using a function provided by the toolbox SPM version 8. Subjects from the sub-population 1 (figure 6) learned to respond in the nogo to win condition and performed worse in the go avoid to losing condition. These findings could suggest that sub-population 1 followed a Pavlovian policy. However, subjects from sub-population 1 also performed worse in the nogo to avoid losing condition compared to sub-population 2, which was not in accordance with a Pavlovian behavior. In order to further investigate whether the behavioral differences between the two groups were instrumental or Pavlovian, we used the Pavlovian parametric estimates (π) of individual fits of the QL+*pav* model. The comparison revealed that there was no significant difference in the Pavlovian parameter between both groups ($p = 0.3321$). This finding suggested that the difference between sub-population 1 and sub-population 2 might not came from the difference in policy. A second hypothesis was that sub-population 1 gathered all the subjects that learned worse in all conditions.

Although, we could not find the source of such a difference, it was clear that the nogo to win condition was inverted in the sub-population 1 comparing to the sub-population 2 (yellow lines in figure 7). This inversion could be the reason why win conditions did not exhibit a significant positive Pearson's correlation in our first analysis. In order to test this hypothesis, we re-computed the Pearson's correlation coefficients, separately, for each sub-population.

The results showed a positive and strong correlation between the go to win and nogo to win conditions in sub-population 2 ($r = 0.947, p < 0.001$). In sub-population 1, the correlation was negative ($r = -.578, p = 0.030$), which is plausible if we consider that subjects followed a more Pavlovian policy in the nogo to win condition.

8 Conclusion

This study aimed to determine whether the PEs used to update old predictions were determined by the value of the action (QL) or by the value of the state (AC). This was achieved by conducting a modified probabilistic Go/NoGo task which orthogonalizes action and valence in 24 healthy subjects.

The behavioral analysis suggested that subjects had correctly learned the go to win, go to avoid losing and nogo to avoid losing conditions, but not the

nogo to win. However, this might be interpreted as an effect of the strong action by valence interaction rather than bad learning. The subjects have learned worse the nogo to win (compared to the nogo to avoid losing condition) and the go to avoid losing conditions (compared to the go to win condition). However, the performance is not disrupted in the same manner in both conditions. In the go to avoid losing condition, subjects tend to perform correctly but they are slower to achieve the correct policy. On the other hand, subjects do not really learn to not respond in the nogo to win condition. This is in line with previous findings which suggested that the learning process was not purely instrumental but it was also affected by a Pavlovian mechanism.

Several Q-learning and Actor-critic models were fitted to the subjects' behavioral data and then compared in order to select the best model. This analysis suggested that the best model was based on a QL framework with a strong, but not persistent bias towards emitting a go choice.

Despite of explaining better the subjects' behavior, the QL was not able to naturally capture the action by valence interaction as well as the Actor-critic model. This was in line with our initial expectations that hypothesized that the learning process tended to be naturally slower in the nogo to win and go to avoid losing conditions when based on an AC framework.

These findings clearly suggested that the subjects' choices came out from a computational mechanism more similar to the QL than to the Actor-critic. In order to further test this hypothesis, we applied a principal component analysis to determine associations among conditions.

The PCA grouped together the conditions which share the same sign of the PEs in the Q-learning framework: the win conditions are linked to positive PEs, the avoid losing conditions are determined by negative PEs and the neutral condition are connected to zero PEs.

PEs are coded by phasic DA, and thus, ultimately, DA is the main driving force of the win and the avoid conditions. Particularly, augmented DA enhances learning from positive PEs (DA bursts), but would impair learning from negative PEs (DA dips). Conversely, reduced DA would enhance learning from negative PEs, but would impair learning from positive PEs. According to this, it would be plausible to expect that the avoid losing and win conditions shared the same component with a negative association between them. Nevertheless, there were evidences which suggested that negative PEs might be conveyed by the dip duration instead of its magnitude. This can explain why win and avoid losing conditions showed different driving forces, and, thus, belonged to orthogonal components.

Furthermore, a clustering analysis showed the presence of two sub-groups which clearly differed in the subjects' performance in the nogo to win condition. In sub-group 1 subjects learned to respond whereas in the sub-group 2 they learned correctly to not respond. This fact could be masking the true correlation between the win conditions. Indeed, the subgroup where subjects have correctly learned to not respond exhibited a strong and positive correlation.

In sum, our findings suggests that healthy humans use the action-values (QL) instead of the state-values (AC) to determine the prediction errors when learning to make a choice. This conclusion corroborates electrophysiological findings in animals, which demonstrated that prediction errors were determined by the action value.

References

[1] R. M. Yerkes and S. Morgulis, "The method of pawlow in animal psychology," *Psychological Bulletin*, vol. 6, pp. 257–273, 1909.

[2] E. L. Thorndike, *Animal intelligence: experimental studies*, E. L. Thorndike, Ed. The Macmillan Company, 1911.

[3] B. F. Skinner, "Two types of conditioned reflex and pseudo type," *Journal of General Psychology*, vol. 12, pp. 66–77, 1935.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*, R. S. Sutton and A. G. Barto, Eds. MIT Press, 1998.

[5] T. V. Maia, "Reinforcement learning, conditioning, and the brain: Successes and challenges," *Cognitive, Affective, Behavioral Neuroscience*, vol. 9, pp. 343–364, 2009.

[6] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593–1599, 1997.

[7] P. R. Montague, P. Dayan, C. Person, and T. J. Sejnowski, "Bee foraging in uncertain environments using predictive hebbian learning," *Nature*, vol. 377, pp. 725–728, 1995.

[8] P. R. Montague, P. Dayan, and T. J. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive hebbian learning," *The Journal of Neuroscience*, vol. 16, pp. 1936–1947, 1996.

[9] J. R. Wickens, 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*, vol. 70, pp. 1–5, 1996.

[10] J. N. J. Reynolds and J. R. Wickens, "Substantia nigra dopmaine regulates synaptic plasticity

and membrane potential fluctuations in the rat neostriatum, *in vivo*," *Neuroscience*, vol. 99, pp. 199–203, 2000.

[11] J. N. J. Reynolds, B. I. Hyland, and J. R. Wickens, "A cellular mechanism of reward-related learning," *Nature*, vol. 413, pp. 67–70, 2001.

[12] J. N. J. Reynolds and J. R. Wickens, "Dopamine-dependent plasticity of corticostriatal synapses," *Neural Networks*, vol. 15, pp. 507–521, 2002.

[13] R. Kawagoe, Y. Takikawa, and O. Hikosaka, "Expectation of reward modulates cognitive signals in the basal ganglia," *Nature Neuroscience*, vol. 1, pp. 411–416, 1998.

[14] ——, "Reward-predicting activity of dopamine and caudate neurons - a possible mechanism of motivational control od saccadic eye movement," *Journal of Neurophysiology*, vol. 91, pp. 1013–1024, 2004.

[15] O. Hirosaka, K. Nakamura, and H. Nakahara, "Basal ganglia orient eyes to reward," *Journal of Neuroscience*, vol. 95, pp. 567–584, 2006.

[16] T. V. Maia and M. J. Frank, "From reinforcement learning models to psychiatric and neurological disorders," *Nature Neuroscience*, vol. 14, pp. 154–162, 2011.

[17] C. R. Gerfen, "Molecular effects of dopamine on striatal-projection pathways," *Trends Neuroscience*, vol. 23, pp. 64–70, 2000.

[18] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, "Representation of action-specific reward values in the striatum," *Science*, vol. 310, pp. 1337–1340, 2005.

[19] A. G. Barto, *Models of information processing in the basal ganglia*. MIT Press, 1994, ch. Adaptive critic and the basal ganglia.

[20] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, 1989.

[21] G. A. Rammery, "On-line q-learning using connectionist systems," Cambridge University Engineering Department, Tech. Rep., 1994.

[22] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman, "Midbrain dopamine neurons encode decisions for future action," *Nature Neuroscience*, vol. 9, pp. 1057–1063, 2006.

[23] M. R. Roesch, D. J. Calu, and G. Schoenbaum, "Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards," *Nature Neuroscience*, vol. 10, pp. 1615–1624, 2007.

[24] M. J. Crockett, L. Clark, and T. W. Robbins, "Reconciling the role of serotonin in behavioral inhibition and aversion: acute tryptophan depletion abolishes punishment-induced inhibition in humans," *The Journal of Neuroscience*, vol. 29, pp. 11 993–11 999, 2009.

[25] M. Guitart-Masip, Q. Huys, L. Fuentemilla, P. Dayan, E. Duzel, and R. Dolan, "Go and no-go learning in reward and punishment: Interactions between affect and effect," *NeuroImage*, vol. 62, pp. 154–166, 2012.

[26] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[27] K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston, "Bayesian model selection for group studies," *NeuroImage*, vol. 46, pp. 1004–1017, 2009.

[28] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly, "By carrot or by stick: Cognitive reinforcement learning in parkinsonism," *Science*, vol. 306, pp. 1940–1943, 2004.

[29] R. B. Rutledge, S. C. Lazzaro, B. Lau, C. E. Myers, M. A. Gluck, and P. W. Glimcher, "Dopaminergic drugs modulate learning rates and perseveration in parkinson's patients in a dynamic foraging task," *The Journal of Neuroscience*, vol. 29, pp. 15 104–15 114, 2009.

[30] T. V. Wiecki and M. J. Frank, "Neurocomputational models of motor and cognitive deficits in parkinson's disease," *Progress in Brain Research*, vol. 183, pp. 275–297, 2010.

[31] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, J. Cohen, Ed. Lawrence Erlbaum Associates, 1988.