NÖRO PSİKİYATRİ ARŞİVİ
ARCHIVES OF NEUROPSYCHIATRY

RESEARCH ARTICLE/ARAŞTIRMA MAKALESİ

Arch Neuropsychiatry 2020
https://doi.org/10.29399/npa.24772

Towards Automaticity in Reinforcement Learning: A Model-Based Functional Magnetic Resonance Imaging Study

Pekiştirmeli Öğrenmede Otomatikleşme: Model Tabanlı Fonksiyonel Manyetik Rezonans Görüntüleme Çalışması

Burak ERDENİZ[1], **John DONE**[2]

[1]Department of Psychology, İzmir University of Economics, İzmir, Turkey
[2]Department of Psychology and Sports Sciences, School of Life and Medical Sciences, University of Hertfordshire, Hatfield, United Kingdom

ABSTRACT

Introduction: Previous studies showed that over the course of learning many neurons in the medial prefrontal cortex adapt their firing rate towards the options with highest predicted value reward but it was showed that during later learning trials the brain switches to a more automatic processing mode governed by the basal ganglia. Based on this evidence, we hypothesized that during the early learning trials the predicted values of chosen options will be coded by a goal directed system in the medial frontal cortex but during the late trials the predicted values will be coded by the habitual learning system in the dorsal striatum.

Methods: In this study, using a 3 Tesla functional magnetic resonance imaging scanner (fMRI), blood oxygen level dependent signal (BOLD) data was collected whilst participants (N=12) performed a reinforcement learning task. The task consisted of instrumental conditioning trials wherein each trial a participant choose one of the two available options in order to win or avoid losing money. In addition to that, depending on the experimental condition, participants received either monetary reward (gain money), monetary penalty (lose money) or neural outcome.

Results: Using model-based analysis for functional magnetic resonance imaging (fMRI) event related designs; region of interest (ROI) analysis was performed to nucleus accumbens, medial frontal cortex, caudate nucleus, putamen and globus pallidus internal and external segments. In order to compare the difference in brain activity for early (goal directed) versus late learning (habitual, automatic) trials, separate ROI analyses were performed for each anatomical sub-region. For the reward condition, we found significant activity in the medial frontal cortex ($p<0.05$) only for early learning trials but activity is shifted to bilateral putamen ($p<0.05$) during later trials. However, for the loss condition no significant activity was found for early trials except globus pallidus internal segment showed a significant activity ($p<0.05$) for later trials.

Conclusion: We found that during reinforcement learning activation in the brain shifted from the medial frontal regions to dorsal regions of the striatum. These findings suggest that there are two separable (early goal directed and late habitual) learning systems in the brain.

Keywords: Predicted value, medial frontal cortex, prediction error, striatum, reinforcement learning

ÖZ

Amaç: Önceki çalışmalar, medyal prefrontal kortekste bulunan birçok nöronun öğrenme süreci boyunca ateşlemelerini tahmin edilen en yüksek değere sahip olan ödüle yönelik yaptığını, ancak daha sonraki denemeler sırasında ateşlemelerini azaltıp, striatum tarafından kontrol edilen daha otomatik bir bilgi işleme tarzına geçtiğini göstermiştir. Mevcut araştırmada önceki kanıtlara dayanarak, öğrenmenin erken dönemlerinde yapılan seçimler için önce öngörülen ödül değerinin medyal frontal korteks tarafından yürütülen hedefe yönelik bir sistem tarafından kodlandığını, fakat ilerleyen denemelerde öngörülen değerin striatum tarafından temsil edildiği hipotezinde bulunulmuştur.

Yöntem: Bu çalışmada, 12 katılımcıdan her biri bir pekiştirmeli öğrenme görevi gerçekleştirirken fonksiyonel manyetik rezonans görüntüsü (fMRG) verisi toplanmıştır. Bu görev, bir katılımcının para kaybetmek veya kazanmaktan kaçınmak için mevcut iki seçenekten birini seçtiği denemelerden oluşmuştur. Buna ek olarak, katılımcılara her bir seçimden sonra para ödülü (para kazanma), para cezası (para kaybetme) veya nötr geri bildirim verilmiştir.

Bulgular: Olayla ilgili fonksiyonel manyetik resonans görüntülüme (fMRG) verileri model tabanlı analiz kullanılarak, olayla ilgili bölgeler olan medyal frontal korteks, kaudat çekirdeği, putamen ve globus pallidus iç ve dış segmentlerine yapılmıştır. Beyin aktivitesindeki farklılığın erken (hedefe yönelik) ve geç öğrenme (alışılmış, otomatik) ile karşılaştırılması amacıyla her anatomik alt bölge için ayrı birer ilgili bölge istatistiksel analizi yapılmıştır. Ödül koşulu için, sadece erken öğrenme denemelerinde medyal frontal kortekste anlamlı aktivite bulunmuş ($p<0,05$), ancak daha sonra bu aktivite iki taraflı putamenlere doğru kaydığı gözlenmiştir ($p<0,05$). Kayıp durumunda ise erken dönem öğrenme denemeleri için anlamlı bir aktivite bulunmamıştır, ancak öğrenme deneyiminin ileri safhalarında globus pallidus iç segmentinde anlamlı bir aktivite görülmüştür ($p<0,05$).

Sonuç: Olayla ilişkili fMRG tasarımı için modele dayalı analizi kullanarak oluşan aktivasyonun öğrenme süresince medyal frontal korteksten dorsal striatuma kaydığı gözlenmiştir. Bu bulgular, erken öğrenme dönemi ve daha sonraki alışkanlık dönemi için iki ayrılabilir sistem olduğunu göstermektedir.

Anahtar Kelimeler: Tahmin edilen değer, medial frontal korteks, tahmin hatası, striatum, pekiştirmeli öğrenme

Cite this article as: Erdeniz B, Done J. Towards Automaticity in Reinforcement Learning: A Model-Based Functional Magnetic Resonance Imaging Study. Arch Neuropsychiatry 2020.

Correspondence Address/Yazışma Adresi: Burak Erdeniz, Department of Psychology, İzmir University of Economics, İzmir, Turkey • **E-mail:** burak.erdeniz@ieu.edu.tr
Received/Geliş Tarihi: 10.06.2019, **Accepted/Kabul Tarihi:** 11.11.2019, **Available Online Date/Çevrimiçi Yayın Tarihi:** 30.01.20200

INTRODUCTION

Over the past century associative learning and more particularly the role of dopamine in associative learning has been studied using mainly two behavioural paradigms: Pavlovian learning and instrumental learning. During instrumental conditioning the outcome (such as food) is contingent on the animal's behaviour. The animal has to perform an action to receive a reward (maximize the amount of food) or to avoid a punishment (minimize the amount of foot shock). Therefore, in instrumental learning the unconditional stimulus becomes a reinforcer to motivate the animal to perform certain behaviours and will give the animal some control over the environment.

In many of the experiments that involve learning stimulus reward associations, learning depends not only on the sequential occurrence of conditional stimuli and the reinforcers but depends on the discrepancy between the actual occurrence and the predicted occurrence of reinforcers (1, 2). Although much evidence has accumulated on the role of dopamine in reward processing in the last 50 years, the dopamine hypothesis of reward has undergone refinement several times (3). These refinements suggest that the specific role of mesolimbic dopamine neurons may be more important for the acquisition of the reward-related behaviours than for subjective responses to rewards (3). A well-established influential theory about the role of dopamine in learning is that of Schultz and colleagues (1). According to Schultz and colleagues this theory is called the reward prediction-error theory and has its roots in the Rescorla-Wagner learning rule (4) and more particularly in the temporal-difference reinforcement-learning model of Sutton and Barto (4). According to reward prediction-error theory, dopamine activations to reward-predicting stimuli occur in almost 80% of dopamine neurons in the substantia nigra and in the ventral tegmental area (1, 2). It was argued that several regions receive this prediction-error signal, including nucleus accumbens, medial frontal cortex, the dorsolateral prefrontal cortex, the amygdala and the organism make predictions about future events (5). Schultz and colleagues proposed that learning occurs by sending back and forth the error-signal between different regions. According to Schultz and Dickinson, it is possible that dopamine neurons may utilize the information about predicted rewards for the control of goal directed behaviour, and they suggested that this information helps to construct reward expectations in the form of predicted values whereas the prediction-error signal generated by dopamine neurons used to update the predicted-values associated with states and actions and these predicted-values might possibly stored in cortical and subcortical regions (1).

The evidence for the prediction error signals in humans comes only recently with the advances in human brain imaging techniques. Functional magnetic resonance imaging (fMRI) studies and their combination with computational models allow researchers to test specific hypotheses about the prediction errors and predicted-values (5). Previous studies showed that reward-predicting stimuli elicit neural activity during instrumental conditioning (6, 7), where the conditional stimuli determine not only the reward predicted by the stimuli but also the action required by the subjects (7). Because the action selection requires motor preparation and movement execution, it has been argued that these types of processes usually comprise of neuronal activity that occurs at the same time as viewing the decision cues (7, 8). It is important to note that sometimes the "predicted-values" might directly relate to specific actions which refer to the future rewards that are expected to be obtained after taking a specific action (e.g., if red light turns on always press the button with the right hand) and can be used interchangeably with action-values (8, 9). Previous studies mainly showed that predicted value activity in the striatum and various cortical regions (10, 11).

More recently, it was showed that much of the human motor behaviour and cognitive processes become automatic after substantial training (i. e., driving a car) (12, 13). In this context over the last ten years our research group examined not only the plausibility of learning by prediction errors (14,15) but we also tested how learning by reward and punishments may affect brain activity (16,17). For example, in a previous study we showed that prefrontal regions together with striatum could sub serve the executive processes involved in early learning and activity in these regions gradual decrease over time (17). In the current study, based on these previous studies, we hypothesized that whether a similar approach can be applied to predicted values of chosen actions. More specifically, we hypothesized that there should be a shift of activation from anterior regions of the brain specifically medial orbito-frontal cortex to posterior regions of the striatum (dorsal striatum; putamen) when the participants learned the predicted values of potential rewards and punishments. Secondly, as previously showed by O'Doherty and colleagues (5), we expected that ventral striatum (nucleus accumbens) should involve in coding the prediction error signal for rewards. Finally, we hypothesized that when the learning is completed (i.e., during late learning trials) there should be no prediction error signal activity in the ventral striatum.

METHODS

Participants

Fifteen healthy normal right-handed volunteers (8 male, 7 female; mean age: 25, range: 22-28) all University students were recruited to the experiment, but only 12 participants (6 male, 6 female) were included in the analysis. Three of the participants were excluded from the analysis, one due to excessive movement inside the scanner (movement greater than 6 mm) and the others due to the loss of behavioural data. Based on participants verbal reports those with prior history of neurological and psychiatric illness were excluded from the study. All participants filled a written informed consent form before fMRI measurements, and all received both written and verbal requests, which outlined the purpose and nature of the study, before the fMRI session. They were debriefed after the experimental session, and paid according to their performance in the task. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Bedfordshire NHS Ethics committee board.

Task

The whole experiment is an event-related fMRI study consisted of 3 sessions, separated by an average of ~2 min. In each session, the color of the stimuli indicated the trial type, except for the neutral trials in which it remained the same for all three sessions (see Figure 2). Within the sessions, each trial was an instrumental learning task involving monetary feedback as a reinforcer. Each trial began with simultaneous presentation of one of three pairs of stimuli (all symbols were letters taken from Agathodaimon font), and each pair of symbols signified the onset of three trial types: Reward (potential monetary gain), Avoidance (potential monetary loss), and Neutral, whose occurrence was fully randomized throughout the experiment. The participant's task in each trial was to choose one of the two symbols by selecting the right or the left key button from the response box. For each pair of stimuli, positions of the symbols (right or left) were also counter balanced within the session. When the trials started, a fixation cross (null event) was shown at the centre of the screen for 0.5 s indicating the start of the trial. This was replaced by the conditional stimulus (two symbols) presented on the screen for 4 s to the left and right of where the cross-had previously been. The participants had to choose which symbol would be rewarded in this 4 s time period. Once the symbol was selected, the chosen symbol was shown by an arrow for 0.5 s followed by the outcome. Between the selected symbol and outcome screens, there was a random inter stimulus interval (ISI) of about average ~2 s for the scanner trigger. The outcome for the participants' choice reward (£1), punishment (-£1) and neutral outcome was shown on the screen for 3 s. The amount of reward and punishment was determined based on the

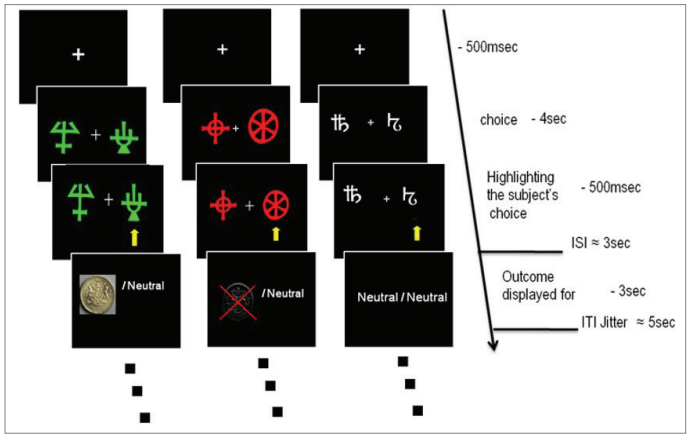

Figure 1. Schematic of the experimental design. Gain trials (green), loss trials (red) and neutral trials (white) were represented in different colors and symbols.

previous studies (18) and recommendations of ethical guidelines (19). No additional payment was made to the participants. When the participants failed to press either button they were instructed at the outcome feedback that they will receive a neutral outcome for the gain pair, or (-£1) for the loss pair. All three trials types were pseudo randomly intermixed throughout the sessions. In the reward trials, when the participants choose the correct symbol (High probability option) then they received monetary reward with 0.8 probability and received neutral feedback with a probability of 0.2. On the other hand, following the choice of incorrect symbol (low probability option), participants received a reward with a probability of 0.2 and neutral outcome with a probability of 0.8. Similarly on the loss trials, if participants chose the correct/optimal symbol (high probability option), they received neutral outcome with 0.8 probability, and a loss outcome with a probability of 0.2, whereas the choice of the incorrect symbol (low probability option) gave a loss of (-£1) with probability 0.8, and a neutral outcome with probability 0.2. On neutral trials, participants always received a neutral outcome independent of the symbol choice. All participants underwent three ~13 min scanning sessions, each consisting of 60 trials (20 trials per condition). Prior to the experiment, participants were instructed that they would be presented with three pairs of stimuli in which the colour of the stimuli would indicate whether it was a reward, loss or neutral trial. They were also instructed that depending on their choices, they would win or lose money or the outcome would be neutral. They were not told which coloured pair of stimuli was associated with a particular type of outcome. All participants were instructed to win as much as possible. Before the experiment, they were told that they could earn a maximum of £30 if they choose the correct response in all trials; otherwise, their earnings would depend on their performance in the experiment.

Functional Magnetic Resonance Image Acquisition

The functional imaging was conducted using 3-Tesla Siemens Magnetom MRI scanner to acquire gradient echo T2* weighted echo-planar (EPI) images with BOLD (Blood Oxygenation Level Dependent Signal) contrast (3x3x3-mm voxel size). Imaging parameters were optimized to minimize signal dropout in medial ventral prefrontal and anterior ventral striatum: we used a tilted acquisition sequence at 30° to the AC-PC line. Each volume was comprised of 36 axial slices of 3-mm thickness and 3-mm in plane resolution with a TR time (repetition time) of 3 s. The flip angle was 90 degrees. T1 weighted structural images (1x1x1-mm voxel size) were also acquired for each participant. Head movement was minimized by head padding.

Pre-processing of Functional Magnetic Image Data

Image analysis was performed using Statistical Parametric Mapping (SPM8) (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, United Kingdom) software. For all participants, the images were realigned according to the first volume in order to correct for motion in the scanner. For all participants, anatomical images were co-registered to functional EPI images, and were normalized to a standard EPI template. Spatial smoothing was applied using a Gaussian kernel with full width half-maximum (FWHM) of 8 mm for each participant's data.

Model-Based Analysis of Functional Imaging Data

The classical correlative paradigm in fMRI simply refers to the manipulation of the independent variables of interest and observing the changes in BOLD response. Model independent paradigms (e.g., epoch analysis) have been useful and are still used by many researchers (13). Nevertheless, it is not good enough to understand value based decision-making and underlying neurocomputational principles (20). According to previous studies most human decisions are usually guided by subjective variables that are not directly observable or controllable by the experimenter (20). These type of variables might depend on a variety of factors such as the subjects' choice history or reward experience and computational models of cognitive processes compute such hidden variables (20, 21). Examples of model-based analysis can also be found for electrophysiological recording studies from behaving monkeys (16). Also these subjective variables differ proportionately to individual differences (e.g., the learning rate). These types of questions have guided researchers to use solutions like model-based techniques. The essential point of the model-based analysis is not whether the brain uses that particular model or not, but most importantly it provides a framework for interpretation and therefore study hidden decision variables and their neural correlates that are critical for learning (22).

The central approach in model-based fMRI is to use the behavioural responses of a participant to estimate the values of the hidden variables of a model over time. In the model-based analysis subjects' behavioural responses were entered into a computational model (i.e, rescorla-wagner learning rule), and the computational model (i.e., implemented by a third party program like matlab) calculates the proxy subjective decision variables (i.e., the prediction error response).

It is important to note that in every computational model there are free parameters such as the learning rate or exploration rate. In the case of reinforcement learning algorithms that need to be calculated with model fitting techniques (e.g., maximum likelihood, or mean least-squares procedure). After all the decision variables are identified and trial to trial parameter values are estimated then they have to be convolved with the hemodynamic function by using parametric modulation and regressed

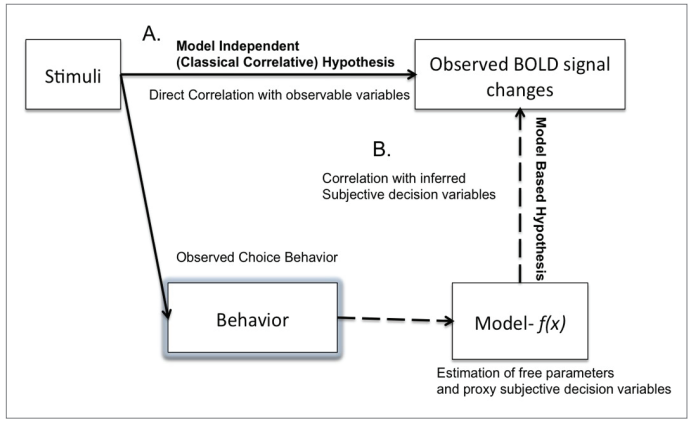

Figure 2. A) In the model Independent classic correlative paradigm, the observable variables are directly correlated against fMRI data. B) In the model-based analysis the hidden (proxy) variables were calculated from the behavioural responses of the participants and then convolved against fMRI data.

against the observed bold signal. The hidden variables are then correlated with fMRI data.

In spite of differences in choice of models, most fMRI studies use similar analysis procedures (i. e., Statistical Parametric Mapping software, https://www.fil.ion.ucl.ac.uk/spm/). In the standard data analysis procedure, images are realigned, spatially normalized to a standard template (for instance MNI or Talairach) and spatially smoothed with a Gaussian kernel. Later, the time series in each session is high-pass filtered to remove potential slow scanner drift or low frequency noise such as heartbeat. After this pre-processing procedure, a statistical linear regression model is fitted to the data. At this point, each trial is represented in the design matrix and the prediction error signal is treated as a parametric modulator to the design matrix. It is important to note that both of the decision variables and free parameter values has to be computed by a second party program (e.g. Matlab, www. mathworks. com). One of the most important issues in calculating the prediction error signal and predicted value is the process of finding the best choice for free model parameters (learning rate and exploration parameter), which were summarized in the following section.

RESULTS

Behavioural Analysis

In order to understand whether there is a statistically significant performance difference between the early (first 10 trials of each condition) and late trials (last 10 trials of each condition), we compared the differences in response times during action selection using a paired t-test. The reason for the first and last half-split is due to participants learning performance where participants reach the asymptotic levels after the first 10 trials (please see Figure 3 and Figure 5). Over the course of the experiment mean response time data for early and late learning trials showed a statistically significant difference between all conditions (reward condition $t_{(11)}$=3.105, p<0.01, two tailed, avoidance condition $t_{(11)}$=4.35, p<0.01, two tailed, neutral condition $t_{(11)}$=5.503, p<0.01, two tailed). These results indicate that during learning participant's response becomes quicker which is an indication of shift towards habit formation or rather automaticity in action selection (Figure 3).

Model Fitting Procedure

During model fitting we estimated each individual's learning rate, α and exploration parameter, β. Here, the updating of the predicted value of the chosen symbol is based on the Rescorla-Wagner prediction error. In order to do that, a softmax action selection rule was used for updating the probability of selected actions. For example, if the participant chose the high probability-option the probability of choosing that option is calculated by the following equation:

$$p(hp) = \frac{e^{\beta Q(hp)}}{e^{\beta Q(hp)} + e^{\beta Q(lp)}} = \frac{1}{1 + e^{-\beta(Q(hp)-Q(lp))}}$$

In the above equation β is the inverse temperature, which inversely relates to the randomness in action selection. For example, high β means higher probability of random action selection (β>0) which is estimated for each participant.

The predicted-values or so-called Q-values inspired from the Q-learning algorithm, high probability (hp) and low probability (lp) were set to 0 at the beginning of each learning session assuming the participants do not have any a priory knowledge about the stimulus values. When the outcome for the particular symbol was presented, the value of the chosen option was updated by the following equation:

$$Q(hp) = Q(hp) + \alpha_t \delta_t$$

In the above equation alpha and delta refers to learning rate and prediction error respectively. To determine the parameters with which the model best fit with the behavioural data of participants' actual choices, we calculated the likelihood function $l(Q|z)$ for each set of parameters ($Q=\alpha, \beta$) with participants actual choices (z) using a custom Matlab program (https://www. mathworks. com). The model fitting procedure is as follows: we first calculate the action values with using all possible combinations of parameter values (incremental search). Then we estimate the probabilities for all possible parameter values for each trial. Then from the probabilities that a participant can select the symbol a in trial i was inserted in the likelihood function. The following equation shows the likelihood function, which is the product of the probabilities in all trials, included in the parameter space, z.

$$l(Q \mid z) = \prod_t p(a,t \mid Q)$$

When we performed a group statistical analysis for the differences in learning rate and exploration parameter, we found that there are no significant differences in learning rate between reward (monetary gain as a positive reinforcer) and avoidance (monetary loss as positive punishment) condition (p>0.05, two tailed), but as expected there is statistically significant difference in the amount of exploration of the

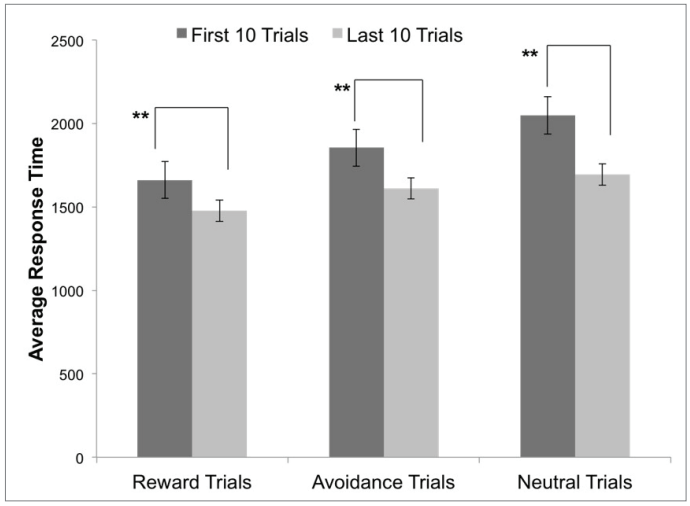

Figure 3. Plot of the reaction times for the three conditions regardless of the outcome received. Participants were significantly slower in the early trials compared to later trials for all conditions. Bars represent standard errors. (**) Represents significance (p<0.05, two tailed). The data represented above belongs to the average of 12 participants.

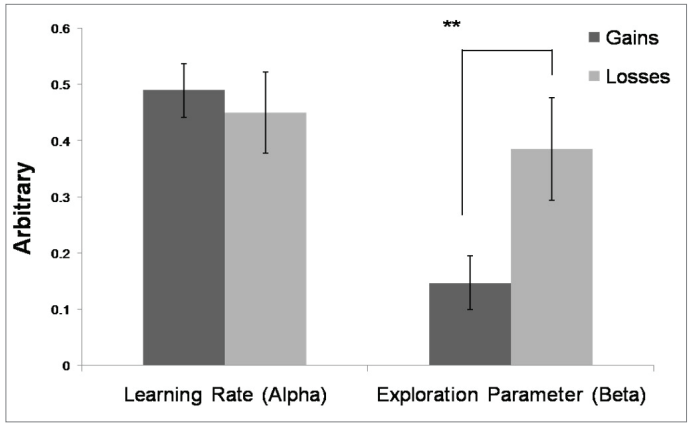

Figure 4. The figure shows that there were no differences in the learning rate of participants between the gain and the loss condition, but there is significant difference in the amount of exploration they perform.

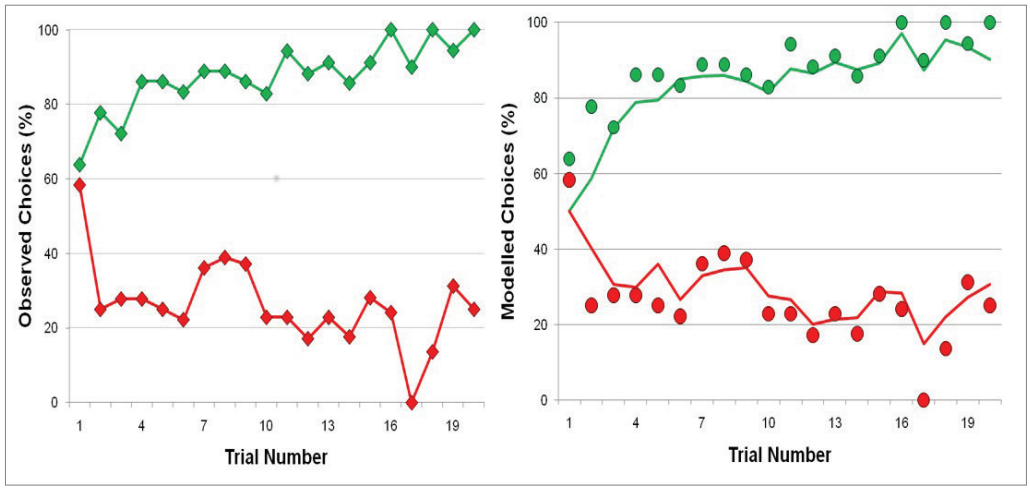

Figure 5. Behavioural model fitting results. Left: observed behavioural choices for reward trials (green) and avoidance trials (red). The learning curves depict, trial by trial, the proportion of participants that chose the high-probability option (symbol associated with a probability of 0.8 of winning £1) for the reward trials (green circles), and the high-probability option (symbol associated with a probability of 0.8 of losing £1) in the avoidance trials (red circles). Right figure: modelled behavioural choices for gain and loss condition. The learning curves represent the probabilities predicted by the computational model. Circles representing observed choices have been left for the purpose of comparison.

other option in the avoidance condition (Figure 4). Based on the higher exploration parameter for avoidance condition we can conclude that participants explore different options more when they were faced with the option that indicate potential losses ($t_{(11)}$=4.3 p<0.05, two tailed).

After estimating each participant's learning rate and exploration parameter, we inserted them into the reinforcement-learning model that was summarized above and calculated the prediction-error (δ) and predicted value of choosing a particular option (Q_{hp}). Also in order to validate how well the reinforcement-learning model fitted with actual choices of participants we looked at the model estimated probabilities of the selected options and actual choices of the participants as can be seen from Figure 5.

The statistical analysis comparing early and late predicted values showed a significant difference for the reward condition ($t_{(11)}$=2.9 p<0.05, two tailed) (Figure 6). The average predicted value of the chosen option (Q-value chosen) in the late trials was significantly greater than the average in the early trials. However the early versus late predicted values for the loss trials were not significantly different from each other ($t_{(11)}$=1.056 p>0.05, two tailed).

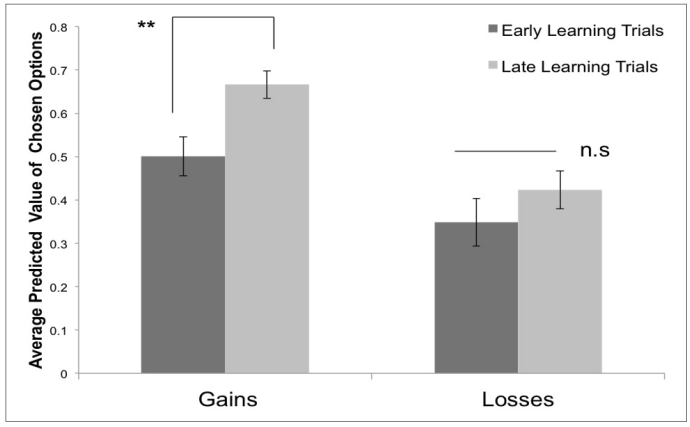

Figure 6. Average changes in the predicted value of chosen option for the early versus late trials for the reward and avoidance condition.

Functional Magnetic Resonance Imaging Results

Using the model parameters described above, we took the trial by-trial predictions of our computational model for the predicted-values and prediction errors for each individual and entered these into first level analysis as a regression model against the fMRI data at the time of stimulus presentation and at the onset of reward/punishment receipt respectively. It is important to note that early (first 10 trials) and late prediction values (last 10 trials) and prediction errors are entered in to the design matrix as separate contrasts. Later on all first level analysis for each individual were carried in to second level group analysis. For the group level analysis, in order to compare the difference in activity for early versus late contrasts separate ROI (region of interest) analyses were performed for each anatomical sub-region using MarsBaR (marsbar.sourceforge.net) tool for statistical parametric mapping software. In the anatomical ROI analysis medial frontal cortex and several sub-compartments of the basal ganglia were used. The specific selection of those ROI regions was based on the previous studies, which showed significant change in BOLD signal for coding reward prediction errors and predicted values of reward outcomes (23). The ROI for the sub-regions of basal ganglia was taken from the BGHAT template (24) and the ROI for the medial orbito frontal cortex is taken from the AAL atlas (25). However, there were no previously defined ROI's in the MNI (Montreal neurological institute) space for nucleus accumbens (NAcc) therefore we have to define NAcc by drawing by hand using MRIcron (http://www.mccauslandcenter.sc.edu/mricro/mricron). The hand drawn NAcc ROI is smoothed with

Table 1. Results of the ROI analysis

	Laterality	t-statistic	p-value
Predicted Value (Early Reward Trials)			
Frontal_Med_ORB_L	L	2.2	0.029
Frontal_Med_ORB_R	R	2.67	0.014
Predicted Value (Late Reward Trials)			
Putamen	L	2.33	0.019
Putamen	R	1.85	0.045
Predicted Value (Late Avoidance Trials)			
Gpi	L	2	0.035
Early Reward Prediction Error			
NAcc	L	2	0.024
NAcc	R	4.43	0.002
Early Avoidance Prediction Error			
Caudate	L	2.86	0.01
Caudate	R	2.78	0.011

Frontal_Med_ORB_L, left medial orbito-frontal cortex; Frontal_Med_ORB_R, right medial orbito-frontal cortex; Gpi, globus pallidus internal segment; NAcc, nucleus accumbens. P, statistical significance.

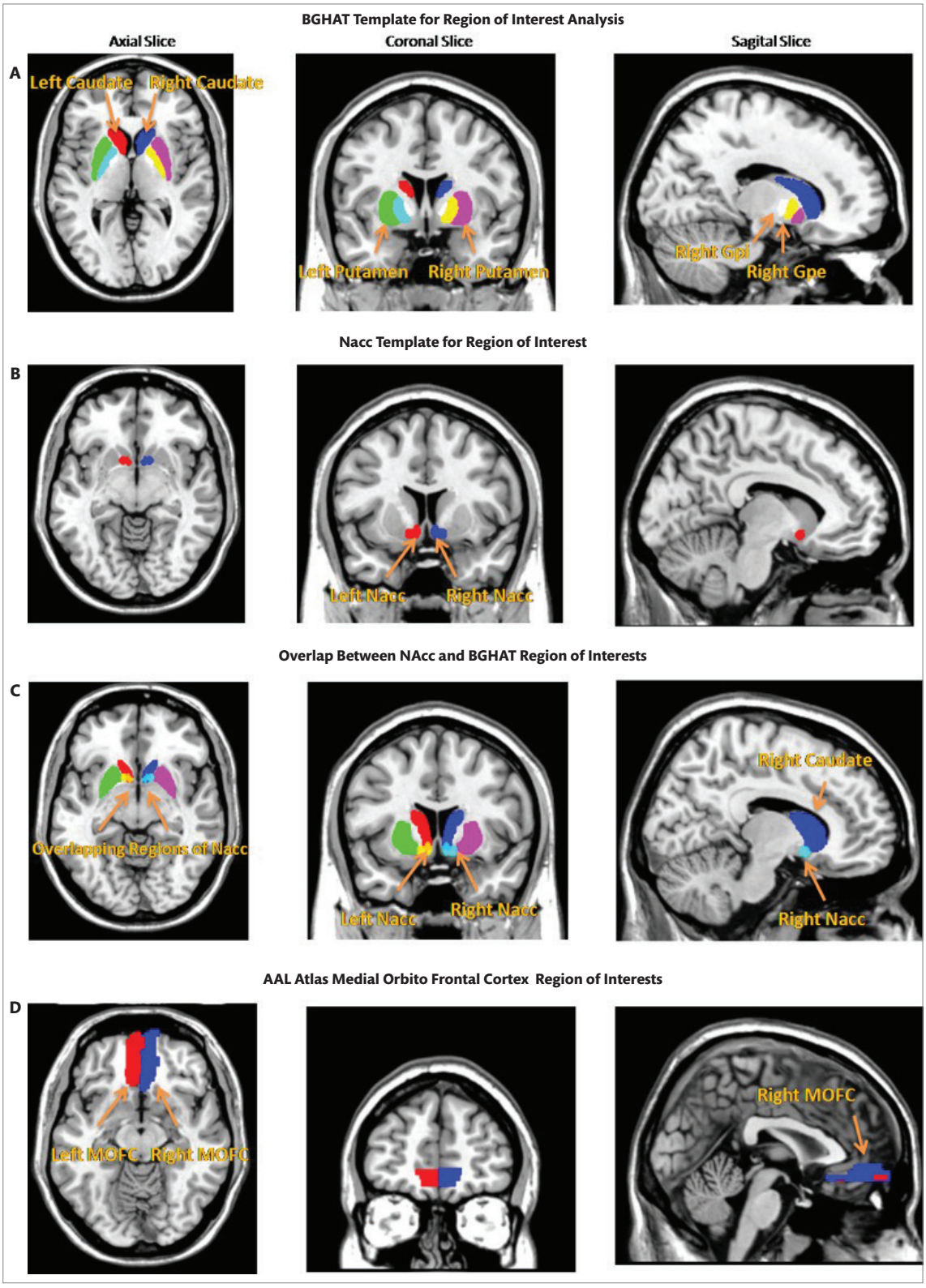

Figure 7. ROI's used in the fMRI analysis. A) BGHAT ROI template for the basal ganglia sub regions. B) Hand drawn ROI for the NAcc. C) Overlapping regions between the ROIs for NAcc, caudate and putamen. D) Medial orbito-frontal cortex ROI based on AAL atlas.

a 3 mm Gaussian kernel and normalized to the Montreal neurological institute (MNI) template. There are 12 regions of interest in total (6 in each hemisphere) and each of these regions was tested separately for 8 contrasts, namely early reward predicted value, late reward predicted-value, early loss predicted value, late loss predicted value, early reward prediction error, late reward prediction error, early loss prediction error, late loss prediction error. This makes a total of 96 test altogether.

The first contrasts that we looked at were the predicted-values of chosen options for reward trials, where we tested each anatomical ROI (including left and right hemispheric regions) separately (see Table 1 for the T-values). The results of the early reward predicted values showed significant positive BOLD change in the medial frontal cortex only and late reward predicted value showed significant change in the bilateral putamen only (Figure 8).

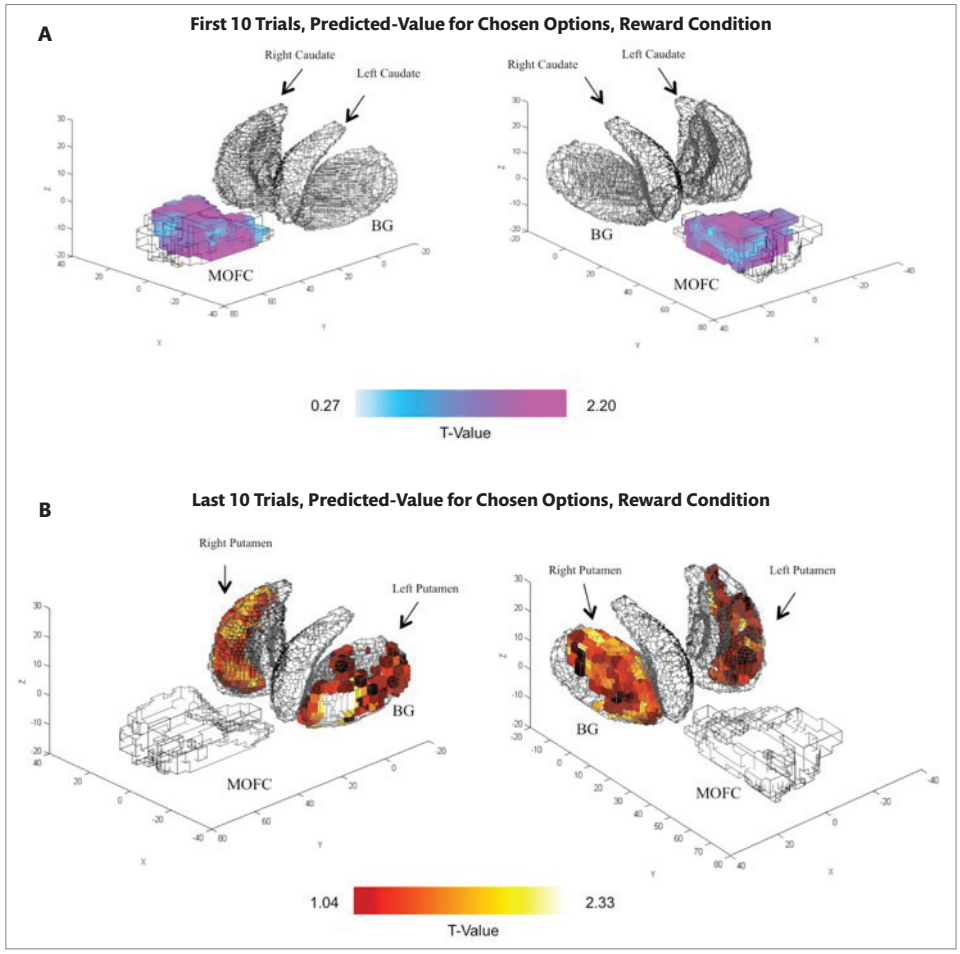

Figure 8. Predicted-values of chosen options during early and late reward trials. a) Activity in the medial frontal cortex correlated with the reward predicted-value in the early learning trials. b) Activity in the right and left putamen is correlated with the reward predicted-value in the late learning trials. The gray mesh frame includes the medial prefrontal cortex ROI (AAL template), the entire basal-ganglia (BGHAT template) and the NAcc ROI. The activations in each voxels have arbitrary dimensions based on multicolor software (www.cns.atr.jp/multi_color_download).

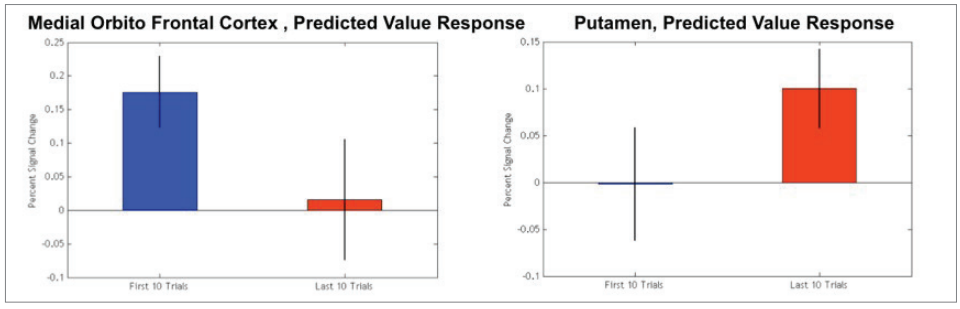

Figure 9. Percent signal changes for predicted-value in the medial-orbito-frontal cortex and putamen. Percent signal changes calculate using the whole ROI region.

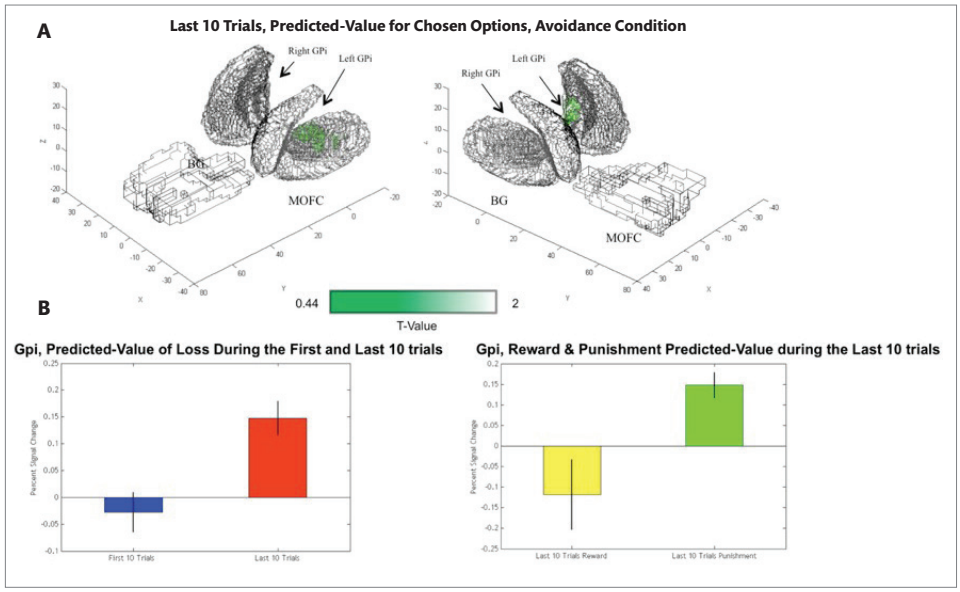

Figure 10. A) BOLD activation for loss predicted-value in the left globus pallidus internal segment. The activity in the ROI overlaid on the mesh frame (gray) that is created by the multicolor software (www.cns.atr.jp/multi_color_download). B) Percent signal change in the left Gpi for the first and the last ten trials. Bottom right. Percent signal change for the last ten trials of reward predicted-value (yellow bar) and loss (green) predicted value.

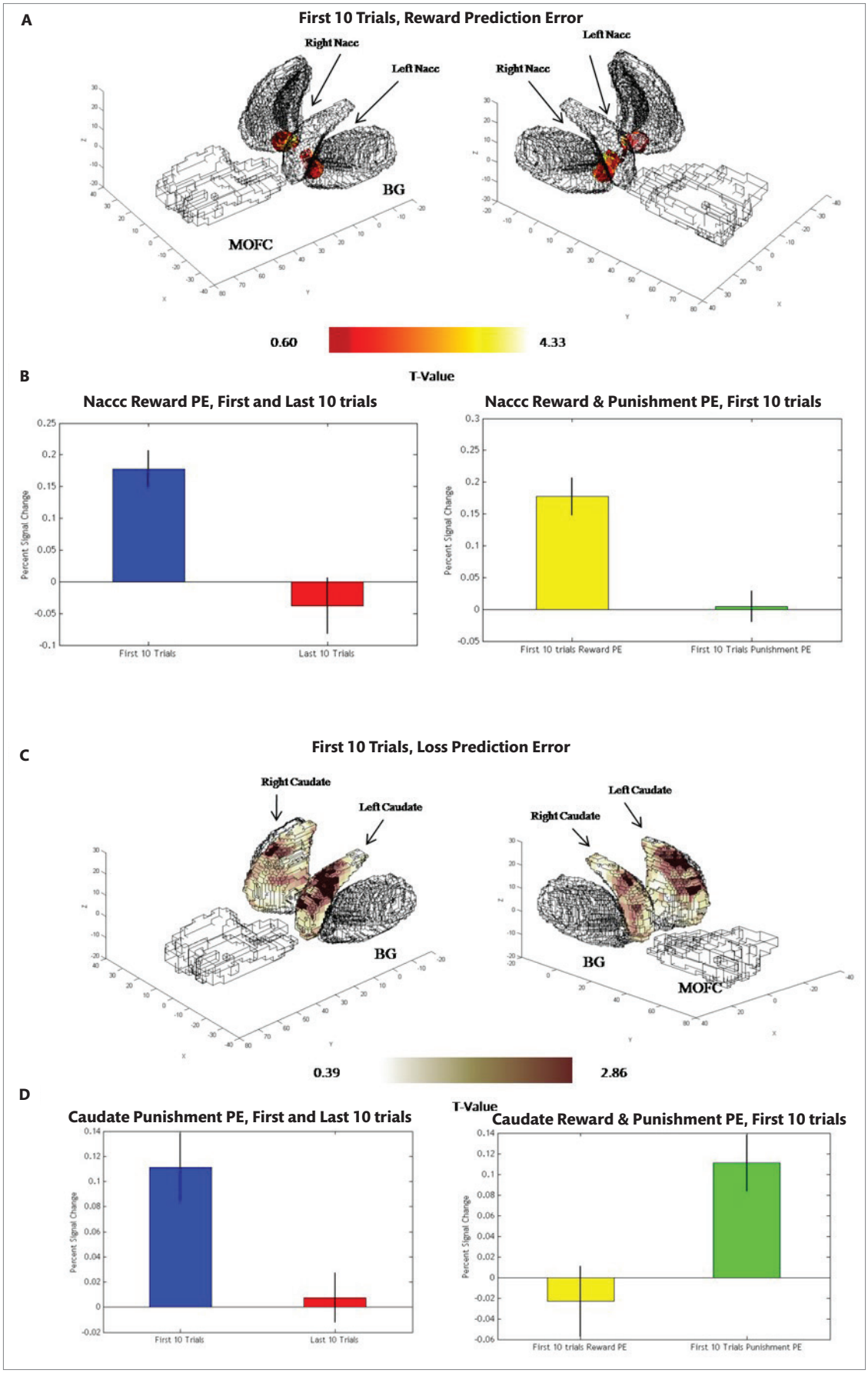

Figure 11. A) Activity in the bilateral NAcc for the reward prediction error during early learning. B) Percent signal change for the first and the last ten trials for the reward prediction error (figure on the left). Percent signal change for the reward and punishment prediction error for the first ten trials (figure on the right) C) Activity in the bilateral caudate nucleus for the loss prediction error during early learning. D) Percent signal change for the loss prediction error in the caudate nucleus for the first and last ten trials of learning.

We also looked at percent signal changes for the predicted value of rewarding stimuli (see Figure 9). We found that the medial frontal cortex is sensitive to reward predicted values early in learning but putamen is sensitive to later in learning

Secondly, we tested for statistically significant changes in signal in these ROIs for early and late loss predicted-values. None of the ROI's showed significant signal change for the early loss predicted-value ($p<0.05$, bonferroni corrected). For the late loss predicted-value we found significant signal change ($p<0.05$, bonferroni corrected) in the left globus pallidus internal segment (Gpi). The comparison of this region with the late reward predicted-value showed that this region was only sensitive to loss predicted values.

Finally, we carried out the same analysis for loss and reward prediction errors. We found that both reward and loss prediction errors produce significant effects only during the first 10 trials and no significant effects were found in any of the ROIs' during late learning trials. For the reward prediction error during early trials, significant activity was found in the bilateral NAcc and for the loss prediction error, we found significant activity in the bilateral caudate nucleus (Figure 11). We also looked for the percent signal change in the NAcc for the loss prediction errors and vice versa for the caudate nucleus for reward prediction error in order to examine the possibility that these regions are specific for loss (Figure 11). We found that the caudate showed negative BOLD signal for the reward prediction error and NAcc showed no percent signal change.

DISCUSSION

Understanding how learning related changes influence brain activity is an important research question (17) and covers a broader research field including the effects of potential gains and losses as well as their opponents (16). In the current study, we identified brain areas responding to changes in predicted-values and prediction errors for early versus late learning trials. We found that during early learning the reward predicted values activate the medial-orbito frontal cortex but later in learning this activity shifts to putamen. We also found that left globus pallidus shows significant activity for the loss predicted-values during late learning trials. In addition to that, we also replicated the well-established findings that showed prediction error signal in the ventral striatum for early learning trials only, whereas loss prediction error activated caudate nucleus and showed no prediction-error related activity during late learning trials.

Recent studies have suggested that there might be more than one type of value signal utilized by the brain (26). It was suggested that the predicted-value signals are involved in the processes of evaluating the anticipated outcome (8). The results of the current study showed significant dissociation between medial-orbito-frontal cortex and putamen supporting the hypothesis that separate brain regions are involved in goal directed and habitual learning. Correspondingly, it has been suggested that the sensori-motor striatum is important in chunking motor patterns in the form of habits and the associative-striatum is important in goal directed learning (26).

Predicted Values in the Striatum

Regarding the findings related to late predicted values in the dorsal striatum many studies suggested that pre movement firing of striatal neurons usually influenced by reward predicting cues (27). For example, earlier studies showed that neurons that fired to initiate movements showed greater excitation when the instruction indicated that the movement was to be rewarded (11). These early studies showed that before the motor actions took place striatal neurons enhanced their firing rate by the information that movement will result in a rewarding outcome (6). This enhancement probably serves to increase the probability of movement in the direction that maximizes reward (e.g., predicted-value of the chosen option or action).

For many researchers the dorsal striatum is the key area for coding predicted-values of options and actions. Perhaps because it has been thought that the main function of the dorsal striatum is related to the preparation and execution of movements (28, 29). More recently Hori and colleagues studied how dorsal-striatal neurons code for action-values by recording from the putamen of monkeys before and after action execution in a go-nogo task (29). They showed that most of the neurons (~50%) in the putamen code for action-values before and after action execution. In another experiment, Pasquereau and colleagues (30) compared action-value (i. e. the action values prior to action execution) and chosen-value (values at the time of action execution) in the putamen and globus pallidus internal segment (Gpi). They showed that in the period of learning the number of action value neurons in the Gpi increased firing and both of the structures influenced by incentive value during the execution of motor responses. For some researchers the increase in the number of neurons that discharge for action value is an indication of automatization during learning (31). Human brain imaging experiments also support the findings on predicted values in the basal ganglia (20). Overall, these studies showed converging evidence for representation of predicted values in the basal ganglia.

Predicted Values in the Cortex

Electrophysiological studies showed predicted-value type of activity in various cortical regions such as the dorso-lateral-prefrontal cortex (DLPFC), parietal cortex, rostral anterior cingulate cortex, and frontal eye fields (FEF) (32). Moreover, by using a computational model of choice behaviour they showed that this activity was highly influenced by the history of actions and rewards.

With the advances in human brain imaging recent studies showed predicted-value and action-value activity in various cortical regions. Some of these studies showed activity in ventro-medial prefrontal cortex and cingulate cortex (33,34). These findings converges with the earlier electrophysiological studies in monkeys (35). Even though this distributed predicted-action-value network is not easy to interpret, one can easily see that parietal cortex and frontal eye fields can only code for effector-specific predicted values. Also, perhaps because of the somatotopic body representations in the striatum, it is possible that the striatum responds to predicted values from all motor modalities in an effector-dependent way.

CONCLUSION

Studies on cortico-striatal anatomy showed that the rostral striatum has connections to orbito frontal cortex via the limbic loop and sensori-motor striatum involving putamen has connections to motor and supplementary motor areas via the motor loop (20). Based on the anatomical connections it is plausible that during early in learning medial-orbito-frontal cortex might be engaged in exploring the response alternatives whereas putamen might be fine-tuning motor movements during later trials.

Finally, although this study specifically focused on the neural correlates of predicted values, it is highly related to computational models of reinforcement learning and raises the question about the performance of alternative computational models. These alternative models will be tested in the future and the reader should be cautious in interpreting the results due to limited sample size.

Ethics Committee Approval: Ethics committee approval was received for this study from Bedfordshire NHS Ethics committee board (Date: 24.06.2008 Decision Number: 06/Q0202/21)

Informed Consent: Written informed consent was obtained from healthy participants who participated in this study.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept- BE, JD; Design- BE, JD; Supervision- JD; Resource- JD; Materials- BE; Data Collection and/or Processing- BE; Analysis and/or Interpretation- BE, JD; Literature Search- BE, JD; Writing- BE, JD; Critical Reviews- BE, JD.

Acknowledgements: We would like to thank Mount Vernon Cancer Centre research team and staff members for their kind help and advice during the course of data collection.

Conflict of Interest: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Etik Komite Onayı: Bu çalışma için etik komite onayı Bedfordshire NHS Etik komite kurulundan alınmıştır (Tarih: 24.06.2008 Karar Numarası: 06 / Q0202 / 21)

Hasta Onamı: Yazılı hasta onamı bu çalışmaya katılan sağlıklı katılımcılardan alınmıştır.

Hakem Değerlendirmesi: Dış Bağımsız.

Yazar Katkıları: Fikir- BE, JD; Tasarım- BE, JD; Denetleme- JD; Kaynaklar- JD Malzeme- BE; Veri Toplanması ve / veya İşlemesi- BE; Analiz ve / veya Yorum- BE, JD; Literatür Taraması- BE, JD; Makale Yazımı- BE, JD; Eleştirel İnceleme- BE, JD.

Teşekkür: Mount Vernon Kanser Merkezi araştırma ekibine ve personeline veri toplama sürecinde yardımları ve önerileri için teşekkür ederiz.

Çıkar Çatışması: Yazarların beyan edecek çıkar çatışması yoktur.

Finansal Destek: Yazarlar bu çalışma için finansal destek almadıklarını beyan etmişlerdir.

REFERENCES

1. Schultz W, Dickinson A. Neuronal coding of prediction errors. Annu Rev Neurosci 2000;23:473–500. [CrossRef]
2. Schultz W. Predictive reward signal of dopamine neurons. J Neurophysiol 1998;80:1–27. [CrossRef]
3. Bromberg-Martin ES, Matsumoto M, Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 2010;68:815–834. [CrossRef]
4. Sutton RS, Barto AG. Reinforcement learning: An introduction 1998 Cambridge, MA. MIT Press. [CrossRef]
5. O'Doherty J, Dayan P, Friston KJ, Critchley HD, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron 2003 38:329–337. [CrossRef]
6. Schultz W, Tremblay L, Hollerman JR. Changes in behavior-related neuronal activity in the striatum during learning. Trends Neurosci 2003;26:321–328. [CrossRef]
7. Niv Y, Schoenbaum G. Dialogues on prediction errors. Trends Cogn Sci 2008;12:265–272. [CrossRef]
8. O'Doherty JP. Contributions of the ventromedial prefrontal cortex to goal-directed action selection. Ann N Y Acad Sci 2011;1239:118–129. [CrossRef]
9. Morita K, Morishima M, Sakai K, Kawaguchi Y. Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. Trends Neurosci 2012;35:457–467. [CrossRef]
10. Kawagoe R, Takikawa Y, Hikosaka O. Expectation of reward modulates cognitive signals in the basal ganglia. Nat Neurosci 1998;1:411–416. [CrossRef]
11. Hassani OK, Cromwell HC, Schultz W. Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. J Neurophysiol 2001;85:2477–2489. [CrossRef]
12. Hélie S, Cousineau D, The cognitive neuroscience of automaticity: Behavioral and brain signatures. Cog Sci 2011;6:25–43. [CrossRef]
13. Poldrack RA, Sabb FW, Foerde K, Tom SM, Asarnow, RF, Bookheimer, SY, Knowlton, BJ. The neural correlates of motor skill automaticity. J Neurosci 2005;25:5356–5364. [CrossRef]
14. Garrison J, Erdeniz B, Done J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci Biobehav Rev 2013;37:1297–1310. [CrossRef]
15. Erdeniz B, Rohe T, Done J, Seidler R. A simple solution for model comparison in bold imaging: the special case of reward prediction error and reward outcomes. Front Neurosci 2013;7:116. [CrossRef]
16. Erdeniz B, Done J. Neural correlates of opponent processes for financial gains and losses. Neurol Sci Neurophysiol 2019;36:69–77. [CrossRef]
17. Erdeniz B, Done, J. Common and Distinct Functional Brain Networks for Intuitive and Deliberate Decision Making. Brain Sci 2019;9:174. [CrossRef]
18. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 2006;442:1042–1045. [CrossRef]
19. Dickert N, Grady C. What's the price of a research subject? Approaches to payment for research participation. N Engl J Med 1999;341:198–203. [CrossRef]
20. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sc 2007;1104:35–53. [CrossRef]
21. Corrado G, Doya K. Understanding neural coding through the model-based analysis of decision making. J Neurosci 2007;27:8178–8180. [CrossRef]
22. Samejima K, Doya K, Ueda Y, Kimura M. Estimating internal variables and parameters of a learning agent by a particle filter. Advances in Neural Information Processing Systems, 2004 MIT Press.
23. Haruno M, Kawato M. Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus action-reward association learning. Neural Network 2006;19:1242–1254. [CrossRef]
24. Prodoehl J, Yu H, Little DM, Abraham I, Vaillancourt DE. Region of interest template for the human basal ganglia: comparing EPI and standardized space approaches. Neuroimage 2008;39:956–965. [CrossRef]
25. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 2002;15:273–289. [CrossRef]
26. Rangel A, Hare T. Neural computations associated with goal-directed choice. Curr Opin Neurobiol 2010;20:262–270. [CrossRef]
27. Balleine BW, Delgado MR, Hikosaka O. The Role of the Dorsal Striatum in Reward and Decision-Making. J Neuroscience 2007;27:8161–8165. [CrossRef]
28. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 2005;22:513–523. [CrossRef]
29. Hori Y, Minamimoto T, Kimura M. Neuronal Encoding of Reward Value and Direction of Actions in the Primate Putamen. J Neurophysiol 2009;102:3530–3543. [CrossRef]
30. Pasquereau B, Nadjar A, Arkadir D, Bezard E, Goillandeau M, Bernard B, Christian Gross E, Boraud T. Shaping of motor responses by incentive values through the basal ganglia. J Neuroscience 2007;27:1176–1183. [CrossRef]
31. Graybiel AM. Habits, rituals, and the evaluative brain. Annu Rev Neurosci 2008;31:359–387. [CrossRef]
32. Sugrue LP, Corrado GS, Newsome WT. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci 2005;6:363–375. [CrossRef]
33. Watanabe M, Hikosaka K, Sakagami M, Shirakawa S. Coding and monitoring of motivational context in the primate prefrontal cortex. J. Neurosci 2002;22:2391–2400. [CrossRef]
34. Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M. Adaptive coding of action values in the human rostral cingulate zone. J Neuroscience 2009;29:7489–7496. [CrossRef]
35. Wallis JD. Neuronal mechanisms in prefrontal cortex underlying adaptive choice behavior. Ann N Y Acad Sci 2007;1121:447–460. [CrossRef]