The timing of exploratory decision-making revealed by single-trial topographic EEG analyses

Athina Tzovara[1,2], Micah M. Murray[1,2,3], Nicolas Bourdaud[4], Ricardo Chavarriaga[4], José del R. Millán[4], Marzia De Lucia[1,2]

1 Electroencephalography Brain Mapping Core, Center for Biomedical Imaging of Lausanne and Geneva, Switzerland

2 Radiology Department, Vaudois University Hospital Center and University of Lausanne, Switzerland

3 Department of Clinical Neurosciences, Vaudois University Hospital Center and University of Lausanne, Switzerland

4 Chair in Non-Invasive Brain-Computer Interface, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Acknowledgements:

The Swiss National Science Foundation provided financial support (grants #K-33K1_122518/1 to MDL and 310030B_133136 to MMM).

Keywords: Decision-making; EEG; single-trial; Exploration-exploitation

Abstract

Decision-making in an uncertain environment is driven by two major needs: exploring the environment to gather information or exploiting acquired knowledge to maximize reward. The neural processes underlying exploratory decision-making have been mainly studied by means of functional magnetic resonance imaging, overlooking any information about the time when decisions are made. Here, we carried out an electroencephalography (EEG) experiment, in order to detect the time when the brain generators responsible for these decisions have been sufficiently activated to lead to the next decision. Our analyses, based on a classification scheme, extract time-unlocked voltage topographies during reward presentation and use them to predict the type of decisions made on the subsequent trial. Classification accuracy, measured as the area under the Receiver Operator's Characteristic curve was on average 0.65 across 7 subjects. Classification accuracy reached a plateau for each of the subjects after ~510ms on average. We speculate that decisions were already made before this critical period, as confirmed by a positive correlation with reaction times across subjects. On an individual subject basis, distributed source estimations were performed on the extracted topographies to statistically evaluate the neural correlates of decision-making. For trials leading to exploration, there was significantly higher activity in dorsolateral prefrontal cortex and the right supramarginal gyrus; areas responsible for modulating behavior under risk and deduction. No area was more active during exploitation. We show for the first time the temporal evolution of differential patterns of brain activation in an exploratory decision-making task on a single-trial basis.

1. Introduction

In various situations humans are faced with the need to make decisions in order to maximize their potential outcome. Often decisions have to be made in an uncertain environment and are therefore driven by the need to either explore the alternative options or to exploit any acquired information. Subjects need to alternate between these two behaviors as information gathering alone does not necessarily lead to the optimal decision, whereas simple exploitation of the acquired knowledge may leave unexplored other options. This kind of behavior has been studied in the context of reinforcement learning theory through the n-armed bandit paradigm (Suton and Barto 1998; Cohen et al., 2007), where subjects need to repeatedly decide among n-different options, each providing a different reward, chosen from a probability distribution.

Neuroimaging evidence based on the n-armed bandit task has highlighted the role of anterior frontopolar cortex and anterior intraparietal sulcus in exploratory decisions (Daw et al., 2006). More generally, the prefrontal cortex and the anterior cingulate cortex have been repeatedly reported to be involved in decision-making under uncertainty (Hsu et al., 2005; Yoshida and Ishii, 2006; Hampton and O'Doherty 2007; Rushworth and Behrens 2008; Seo et al., 2009). Discrimination between exploratory/exploitatory decisions has also been documented using alpha and beta band EEG activity (Bourdaud et al., 2008).

However, the temporal aspects of decision-making still remain under-explored. In the present study we aim at identifying how early in time the relevant generators start differentiating their responses in order to eventually lead to an exploratory or exploitatory decision. It is known that EEG responses start differentiating according to

the subjects' decisions already from the presentation of reward, at an average across trials and subjects level (Cohen and Ranganath 2007). Modulations of the EEG responses following reward presentation are also present at a single-trial level (Philiastides et al., 2010), allowing to discriminate between switch/stay decisions, although the temporal aspects of this discrimination are not yet explored.

In the present study, in order to investigate fine-grained temporal information, we carried out an EEG experiment while subjects were facing the 4-armed bandit problem with four classes (Daw et al., 2006; Bourdaud et al., 2008). In such a high-level cognitive task, inter-subject variability cannot be neglected as individual subjects employ different strategies (Daw et al., 2006), an effect also linked to genetic polymorphisms (Frank et al., 2009).

We therefore carried out analyses at the single-subject level, using a classification scheme, which allows to discover the neural correlates underlying decision-making that can best predict subjects' behavior (see Hampton and O'Doherty, 2007 and Bourdaud et al., 2008, for similar approaches based on functional magnetic resonance imaging –fMRI- and EEG, respectively). The main difference here is that prediction is not the goal of the study per se as in Bourdaud et al., 2008, but rather a strategy for evaluating statistically when enough information is available for accurately classifying future decisions, as measured by EEG. Without making explicit assumptions about the neural underpinning of decision-making, we consider voltage topographies that best discriminate exploratory and exploitatory behaviours in a time-unlocked manner.

Classification based on voltage topographies has been reported in lower-level tasks in the visual and auditory domains (De Lucia et al., 2007; Murray et al., 2009; Tzovara et al., 2011). Here, we show in a more challenging context that EEG topographies can

accurately predict behaviour with the advantage of being neuropysiologically interpretable: any change in them is the result of a change in the underlying brain generators.

2. Materials and Methods

2.1 Experimental paradigm

2.1.1 Participants

Seven healthy individuals (2 females), aged from 25 to 27 years (mean age 26.4 years), participated. Data from these individuals have been previously published in an investigation on the role of EEG oscillatory activity on single electrodes during the exploration – exploitation task (Bourdaud et al. 2008). In the present study we further analyze the temporal aspects of these data, in association to reward evaluation.

2.1.2 Procedure and Task

The experimental protocol was adapted from a similar fMRI study (Daw et al. 2006). Participants were sitting in front of a computer screen where four squares were displayed representing four slot machines (Figure 1a), where each machine corresponds to a bandit arm. They were instructed to fixate on a red dot at the center of the screen to reduce ocular artifacts. On each trial participants had to choose one machine by pressing a key with their index or middle finger on the corresponding hand (left hand for machines 1 and 3, and right hand for machines 2 and 4). The payoff of the selected machine was displayed one second after the key press and remained on display for another second, followed by the beginning of a new trial. Participants were asked to

select the machines so as to maximize their total gain (i.e., sum of individual payoffs) over a session of 400 trials. Three sessions were recorded for each participant.

The payoff of each machine, a numerical value between 0 and 100, was drawn from a Gaussian distribution whose mean changed slowly across the experiment. Before the experiment, nine random but common across participants examples of the payoff evolution for all the machines were shown to each of them (for such an example see Figure 1b). Participants, knowing that the machines' payoffs were not static, had to regularly update their knowledge about them and were therefore encouraged to explore.

2.1.3 EEG acquisition

Continuous 64-channel EEG was acquired through a Biosemi Active II system with a sampling rate of 2048 Hz and was referenced to the CMS-DRL ground, which functions as a feedback loop driving the average potential across the electrode montage to the amplifier zero. EEG recordings were not performed inside a faraday cage, so as to ease reproducibility of any findings for possible future online applications in real-life conditions. The acquired signal was filtered offline by an eighth-order low-pass Chebyshev Type I filter with a cutoff frequency of 205 Hz and down-sampled to 512 Hz. The filters were applied in both the forward and reverse directions to remove all phase distortion, effectively doubling the filter order. In addition, electrooculogram was recorded using two electrodes located below and at the outer canthus of the right eye.

2.1.4 Preprocessing

Trials were extracted with respect to the display of payoff, spanning 100 ms before the display and 780 ms post-stimulus onset (Figure 1a, red thick line). Trials with blinks or

eye movements were rejected off-line. An artifact criterion of ±70 μV was applied at all electrodes, and each EEG epoch was also visually evaluated. Data from noisy electrodes from each subject and condition were interpolated using three-dimensional splines (Perrin et al., 1987). Each subject's data were 40 Hz low-pass filtered. No baseline correction was applied. Data at all time-instances were normalized by their instantaneous Global Field Power (GFP; Lehmann and Skrandies, 1980; Koenig and Melie-García, 2010) to eliminate any strength influence.

2.1.5 Behavioral model

A behavioral model is required to label each trial as corresponding to either an exploratory or exploitative decision. Here, we use the same behavioral model proposed in Daw et al. 2006 and Bourdaud et al. 2008, involving two steps for every trial: First, it provides an estimation of the payoff that users expect to receive from each machine and in a second step it estimates which machine subjects are supposed to choose. In the latter step each subject is considered separately in order to account for inter subject variability.

Specifically, for the payoff tracking we use a Bayesian linear Gaussian rule (Kalman filter) whose parameters are computed using the available data from all subjects, while for the machine selection we use a softmax rule, separately applied for each of the subjects. In both steps, we estimate the required parameters by maximizing the model likelihood with respect to the subjects' choices.

In order to label the subjects' decisions as exploration or exploitation, we use the estimated payoffs of all the machines. When subjects choose the machine corresponding to the highest estimated payoff, their decision is labeled as exploitation. However, when

they choose the machine that did not correspond to the one with the highest estimated payoff the decision is labeled as exploration, but only if this reward differs from the highest one by a threshold value (set to 4). Only exploratory trials corresponding to a machine change and exploitatory corresponding to a machine stay are kept for further analyses, in a similar way as was applied in another EEG study using the same task and subjects (Bourdaud et al., 2008). We used here exactly the same behavioral model in order to obtain comparable results.

In the following, we use EEG activity during reward evaluation of trial n-1 to predict the subject's choice in trial n. For that reason, we split the trials in conditions based on the label of the following decision. For example, trials referring to the exploration condition include the period of reward evaluation from trials preceding the actual exploratory decision. Overall the two categories (exploration and exploitation) were highly unbalanced and in order to avoid overfitting one of the two, we consider the same number of trials for both conditions by randomly selecting exploiting trials so as to cover the whole time-course of the experiment.

2.2 Average ERP analysis

We first examine whether there are any time-locked temporal periods, common across subjects, during which the relevant neural generators differentiate their responses between exploration and exploitation. To this aim, we compute event-related potentials (ERPs) by averaging peri-stimulus epochs. We then calculate the global dissimilarity (Lehmann and Skrandies, 1980), time-point by time-point, between exploratory and exploitatory trials. Global Dissimilarity computes configuration differences between two

electric fields, independent of their strength and is statistically analyzed by applying a Monte Carlo bootstrapping analysis procedure, colloquially known as topographic ANOVA (TANOVA; Murray et al., 2008). Electric field changes forcibly follow changes in the underlying brain generators, so the TANOVA analysis is a way of determining whether and when the generators of the two types of decisions differ statistically.

Moreover, we examine the relevance of the temporal intervals identified from the TANOVA for predicting the subjects' decisions on a trial-by-trial level. We use an analogous approach to our single-trial analysis (that will be described bellow), by computing the spatial correlation of the voltage topographies observed at the average ERPs for exploration and exploitation with the topographies observed at the same latency at the single-trial level. Each trial is classified as belonging to exploitation if across the time interval identified by the TANOVA its topographies correlate more with the topographies observed on the exploitation condition; otherwise it is classified as exploration.

2.3 Single-trial analysis

Our first goal through the single-trial analysis is to identify those features of the EEG signal that best account for the subjects' decisions on a trial-by-trial basis. For this purpose we use a classification scheme based on single-trial voltage topographies, requiring minimal *a priori* assumptions (De Lucia et al., 2007; Murray et al., 2009; Tzovara et al., 2011). We hypothesize that there exists at least one set of underlying generators (or voltage topographies) per experimental condition (i.e. exploration or

exploitation) responsible for the subsequent decision. However, although there could exist more than one set of generators, our analysis is refined so as to detect exactly one.

We develop an algorithm that extracts the two voltage topographies that best predict the subjects' decision based on modeling their statistical distribution of the entire set of single trials of each of the two conditions (see *2.3.1Model Estimation*). Due to the nature of decision-making phenomena, these discriminant topographies are rather unlikely to be detected locked in time across trials, therefore our algorithm operates independently of the time point at which a given topography occurs. We select these voltage topographies by splitting the entire dataset for each subject in ten subsets of single-trials. We carry out voltage topography estimation using nine splits of the data (on average 113 trials per subject; Training dataset), and test their prediction accuracy on the split that was left aside (on average 12 trials per subject; Test dataset). This procedure provides an estimation of the best discriminant topographies and an indication of the average prediction accuracy based on the test datasets. Finally, we further validate the prediction algorithm on a validation dataset. Trials used for the validation were kept separate from those used for estimating the two discriminant voltage topographies (on average 12 trials per subject). This validation dataset provides an indication of whether it is possible to generalize our findings not only on the test trials, but also on completely new data with similar levels of accuracy.

At a second stage we aim at detecting, for each subject separately, the time-period along which the underlying generators have been sufficiently activated to lead to a correct prediction of the subsequent decision, across trials. This analysis gives an indication of the time-period by which the decisions have already been made across trials.

2.3.1 Model estimation

The first step in the single-trial analysis comprises a modeling of the ensemble of topographies in the training dataset, for each experimental condition separately. All available topographies are pooled together, without taking into account their temporal order and are modeled using a Mixture of Gaussians Model (GMM) in a 64-dimensional space (since EEG was recorded using 64 electrodes). The technical details of the GMM implementation have been reported elsewhere (De Lucia et al., 2007; Murray et al., 2009; Tzovara et al., 2011). Through this modeling procedure we can cluster all the recorded topographies in Q Gaussians in total. The mean of each Gaussian is a topography itself. Therefore, using the Gaussians we can extract a few representative topographies for the whole dataset. In the following, we will refer to the means of the Gaussians as template maps. The modeling of the data is performed separately for the two conditions, and we therefore obtain one set of template maps per experimental condition. We make an *a priori* hypothesis about the total number of Gaussians in the model and we then optimize it so as to achieve the best prediction accuracy (a point to which we will return below).

In order to assess the degree to which each topography is represented by the template maps within one model we use the posterior probabilities:

$$P(c_k|\boldsymbol{m}) = \frac{P(\boldsymbol{m}|c_k) \cdot p_k}{p(\boldsymbol{m})}$$

(1)

Where $\boldsymbol{m}$ refers to a recorded topography, c_k to a particular template within the GMM, $p(\boldsymbol{m})$ to the unconditional density function and p_k to the prior probability of the Gaussian k.

Our goal is to identify the pair of template maps that best discriminates between exploratory and exploitatory decisions. For this purpose, we use the Bayes Factor (Raftery 1995):

$$BF_{t,h} = \frac{P(c_{exploit_k}|\boldsymbol{m_{t,h}})}{P(c_{explor_{k'}}|\boldsymbol{m_{t,h}})}$$

(2)

where $BF_{t,h}$ [1] is the Bayes Factor at latency h, for the t-th trial, and $\boldsymbol{m_{t,h}}$ is the observed topography of the t-th trial, at latency h. $c_{exploit_k}$ and $c_{explor_{k'}}$ refer to the k-th Gaussian (or template map) within the GMM generated for exploitation and to the k'-th Gaussian within the model for exploration condition. k and k' range from 1 up to the total number of Gaussians in each of the models. The Bayes Factor is computed in the same way for topographies belonging to both conditions (see Figure 2 for an example).

For all possible combinations of the template maps in the two models, we compute the Bayes Factor at every latency of every trial. Thus, we obtain a measure of the confidence with which we can assign a specific observation to the template map of the model for exploration or to the one for exploitation. A value of Bayes Factor at a specific latency and trial greater that 1 suggests that the topography recorded at that latency/trial is more likely to be represented by the template map of exploitation, and a value lower than 1 by exploration (see Equation 2).

As discriminant function we consider the average of the $BF_{t,h}$ values along the trial:

$$DF_t = \frac{1}{L}\sum_{h=1}^{L}(BF_{t,h})$$

[1] For simplicity we do not include k and k' in the notation of the left side of the equation 2.

(3)

where L is the length of the trial in time-points. DF_t provides information about the relative degree of presence in time of two topographic maps, no matter at which specific latency within the trial they occur. At this stage the discriminant function is only computed for the overall trial, (for L equal to the trial's length). However, at a later stage we compute an analog of this function for varying values of L, in order to investigate the temporal aspects of decision-making (see *2.3.3 Relevant time-periods for decision-making*).

Our first goal is to identify two template maps, one that reflects the neural correlates of exploratory decisions and one of exploitatory. For this purpose we select within the computed GMMs those two maps that provide the highest levels of discrimination in the training dataset, using DF_t as a discrimination function. The selection is done in the training dataset and then confirmed in the testing dataset, by only using those two maps for performing classification (see below).

2.3.2 Model selection and Classification accuracy based on the overall trial

In order to classify new trials from the test dataset, we compute their DF_t for the two maps that have been identified from the training. In general, we measure classification performance as the area under the Receiver Operator Characteristic Curve (AUC; Green and Swets, 1966). The selection of the maps and the classification are repeated ten times, for every split of the data and the final values of AUC reported are the average values across the ten splits.

To define chance levels, we randomly shuffle the true labels of the test trials and then perform classification. This is repeated 100 times and chance levels are defined as the average AUC over these randomizations.

We remind the reader that our algorithm requires an initial assumption about the total number of Gaussians in the models. To select this value we generate multiple models for each condition, with varying total number of Gaussians, ranging from three up to eleven. The whole procedure described above is repeated for every possible combination of models between the two conditions, providing us different values of classification accuracy. Finally, we select the pair of models that maximizes the mean AUC across the ten splits of the data.

Finally, in order to obtain a more realistic measure of the performance of our method we perform one class classification on completely new trials, not used in any point (training/testing) so far. Due to the limited amount of trials for exploration we were obliged to use all of them for training the models. Therefore, the validation dataset consists only of exploitative trials. Using the DF_t of the trials of the training datasets, we define the optimum threshold for discriminating between the two conditions for every data split. We compute the average ratio of true positives in the validation and in the test dataset and we compare the two for estimating a general classification performance.

2.3.3 Relevant time-periods for decision-making

Using the same number of Gaussians that has been selected as explained above, we further investigate the temporal behaviour of the extracted topographies. For each trial, we consider the expanding average of the Bayes Factor (ABF; Equation 4):

$$ABF_{t,h^*} = \frac{1}{h^*}\sum_{i=1}^{h^*} BF_{t,i}$$

(4)

where t refers to the t-th trial and h^* to a generic latency within the trial. We compute ABF_{t,h^*} for every time-point h^* within a trial, or the average of the Bayes Factor from the beginning (time-point 1) up to h^*, across trials. The ABF value is the equivalent of DF_t (Equation 3), but computed only up to a time-point h^* , with $h^* \leq L$.

Based on the ABF, we now compute the classification time-point by time-point, for all possible values of h^* (with $1 \leq h^* \leq L$) and obtain, for each subject separately, the time-course of AUC values. This allows identifying, in a data-driven way, the time-periods that are relevant for the subjects' decisions while taking full advantage of the fine temporal resolution of the EEG signal.

This expanding average, ABF, is used as a way to examine the temporal patterns of discriminatory activity between exploration and exploitation at the single-subject level. This is an alternative to similar approaches employing fixed sliding windows for finding the period of interest (Philliastides et al., 2010), with the advantage that there is no bias induced by selecting the length of the window.

As will be shown in the Results section, the typical temporal behaviour of the ABF reaches a plateau after a certain time period. After this plateau is reached, ABF remains relatively stable, although it is computed over a longer period, and consequently the classification accuracy does not statistically improve even when considering more time points.

In order to quantify that period of stability, starting from the end of the trial we move back in time, time-point by time-point. We assess for which time-point h* the difference between the AUC value at the end of the trial and the AUC value at a latency h* does not differ significantly from zero.

2.3.4 Time-locked discrimination

As an alternative to the abovementioned procedure, we also perform classification at every time-point. This is done in order to assess whether the extracted topographies start to differ at some specific, time-locked point in time (but potentially varying across subjects). Using the already-selected models and the selected template maps within them, we compute for every time-point the Bayes Factor for all trials, time-locked and use it as discrimination function.

2.3.5 Source estimations

Intracranial sources are estimated using a distributed linear inverse solution and Low Resolution Electromagnetic Tomography (LORETA) regularization approach (Pascal-Marqui et al., 1994, Michel et al., 2004). The sources' current distribution is calculated in a discrete grid of 3005 solution points, regularly distributed within the gray matter of the cerebral cortex and limbic structures of the average brain provided by the Montreal Neurological Institute (MNI). The source estimations are performed on the average of the template maps that were extracted from all subjects. Paired t-tests are calculated at each solution point using the variance across subjects. Only points with values of $p < 0.05$ ($t_{(6)}>2.61$) and clusters of at least 9 contiguous nodes are considered significant. This spatial criterion is determined using the AlphaSim program (available at

http://afni.nimh.nih.gov). The 10,000 Monte Carlo permutations performed on our current distribution matrix revealed a false positive probability of <0.05 for a cluster greater than 9 nodes. The results of the source estimations are rendered on the MNI brain with the locations named using the convention of Talairach and Tournoux (1988).

2.3.6 Disentangling exploratory decision-making from other confounding factors

Due to the nature of the task, several factors can affect the subjects' decisions and influence their behavior. It is therefore important to disentangle decision-making from reward evaluation (in terms of wins/losses) and machine switching.

More specifically, we expect that the subjects' decisions on trial n are influenced by the received reward on trial n-1. It is not clear however whether our results are truly based on a prediction of the subsequent decision or are a reflection of the neural processes of reward evaluation. Using the same training phase and the same models as before, we split the test trials in terms of wins and losses, with respect to the reward prediction error. In general, those trials in which the received reward is lower than the expected one (losses) are more likely to lead to exploration and those where the received reward is higher than the expected (wins) to exploitation (see also *3.1Behavior*). Therefore, we treat those trials accordingly and perform classification to examine whether the extracted template maps can also account for differences between wins and losses.

Moreover, we assessed the influence of machine switching in our results. Because of the way we computed our behavioral models, exploration forcibly corresponds to a machine switch and exploitation to a machine stay, for all the trials we used for training and testing the GMMs. We now only include trials corresponding to switch and stay that were never used before during training and testing. For defining validation trials

corresponding to switch and stay, we dropped the constraint that forced all our exploitation trials to be stays (see *2.1.5 Behavioral model*). Consequently, the 'switch' trials were part of those exploitation trials which were previously labeled as unknown. We use this validation dataset to assess to which extent we can generalize our models also to these specific categories of trials and to examine the extent to which machine switching is influencing our results. Staying with the same machine is inherent to our definition of exploitation, so we expect to obtain high sensitivity but low specificity for this case.

3. Results

3.1 Behavior

Overall, subjects were more likely to exploit than to explore: 61% of the trials were labelled by the behavioural model as exploitation and 13% exploration, the rest were unknown. The validity of the behavioural model employed here has already been demonstrated (Bourdaud et al., 2008) by comparing its results with the actual statistical parameters of the machines.

There was no significant difference between the subjects' reaction times between exploratory and exploitatory decisions (paired t-test; $t(6) = 2.17$; $p = 0.07$), in accordance to what has been reported in similar tasks (Jepma and Nieuwenhuis, 2011).

Through the behavioural model we extracted the reward that the subjects expected to receive for each machine. When subjects experienced a loss (i.e. the actual reward was lower than the expected one) they were most likely to explore on the next trial

(probability across subjects: 0.75 versus 0.16 for exploitation) and when they experienced a win (i.e. the actual reward was higher than the expected one) they were most likely to exploit (probability: 0.84 versus 0.25 for exploring instead).

3.2 Average ERPs

Figure 3 displays the group-average ERPs from a central midline electrode (Cz) for exploitation (blue line) and exploration (green line). ERPs refer to reward evaluation and are time-locked to the appearance on the screen of the payoff of the selected machine (time zero). The average waveform exhibits characteristic features of reward evaluation, the so called feedback-related negativity peaking at 135ms post-stimulus, with a baseline-to-peak amplitude of -0.6 and -0.68μV for exploitation and exploration, respectively. Another component, with a positive amplitude over Cz (2.5 and 2.9 μV for exploitation and exploration respectively), peaks at 225ms, corresponding presumably to a P300. At the same latencies the average scalp topographies are also displayed (Figure 4, blue frame for exploitation and green for exploration). Both components have been reported to be relevant with the evaluation of the displayed reward (Frank et al., 2005; Wu and Zhou 2009).

The TANOVA analysis was performed on an average, across subjects level. It revealed three distinct periods of topographic differences between exploration and exploitation: the first starting at 221ms and lasting up to 252ms, the second at 283-299ms and the third at 676 – 713ms, post-stimulus onset (Figure 3, red periods).

We further tested the relevance of these three intervals for the trial-by-trial decision-making. Only those intervals that were found in at least 8 out of ten training datasets are reported here. The accuracy of classification was computed for each of them separately,

according to the spatial correlation of single-trials with the voltage topographies observed on the average level. Across the ten splits of the data and across the seven subjects, we obtain average AUC values of 0.55±0.02, 0.56±0.03 and 0.55±0.02 (±s.e.m.), for each of the identified intervals, respectively. The low values of prediction accuracy demonstrate that even though topographic differences exist on an across subjects level, neither the identified intervals, nor the average ERP topographies carry enough information for predicting individual subjects' decision at the single-trial level, possibly due to crucial single-trial information that is lost when averaging. We speculate that these, relatively early, time-locked periods of difference are more specific to reward evaluation than to decision-making.

3.3 Time-unlocked prediction

The total number of template maps in the optimal GMMs was in the range of three to six for exploitation and three to eleven for exploration, across subjects. Within all the maps of the GMMs we extracted one template per condition (Figure 2a, for one exemplar subject) and computed the BF across trials (Figure 2b).

Classification was based on the average of the BF across the whole trial (Equation 3). The average AUC values across the ten splits of the data are shown for each subject in Figure 4 (light gray bars). On average across subjects the AUC was 0.65, ranging from 0.55±0.01 (±standard error across ten splits of the data) for the worst in terms of AUC performing subject (S6 in Figure 4) and up to 0.75±0.08 for the best (S4 in the same Figure). Importantly, the AUC was significantly higher than chance levels for six out of seven subjects (Figure 4, all but S5; t-test; $t(9) > 3.2$; $p<0.05$).

The ratio of true positives, on average across subjects and shuffles was 63±2% for exploitation trials of the test datasets and 61±2% for trials belonging to the validation datasets. The two accuracies did not differ significantly across subjects (paired t-test, $t(6) = 0.99$; $p = 0.36$) showing that our models can be generalized with comparable accuracy to new trials.

3.4 Relevant time-periods for decision-making

Based on the expanding average version of the ABF, we computed the classification along the trial (Figure 5). In general, the AUC vales increase over time (Figure 5, solid red lines), as we consider more evidence in the ABF. Note that the AUC values reported previously (3.3 Time-unlocked prediction; Figure 4), correspond to what was obtained at the end of the AUC time-courses, when the classification is based on the overall trial.

We observed that after a certain time-period, the AUC values reach a plateau and remain relatively stable, even though they were computed over a larger amount of data, possibly because the subjects' decisions across trials have already been made. To quantify this time-period, we tested at which time-point h* the AUC values drop significantly across the ten splits of the data, when compared to the values we get using the whole trials (t-test, $p<=0.05$). This plateau was detected for every subject separately and on average it occurred at 508ms post-stimulus onset (Figure 5, gray vertical lines), or ~880ms before the subsequent button press. It is worth noting that this plateau is estimated in statistical terms although the absolute value of AUC may still increase for some subjects.

We assume that these plateaus show when the decisions have already been made across trials, for each subject separately. Interestingly, the onset of these plateaus for each subject was significantly correlated with the subjects' average reaction times on the task (Figure 6–; Spearman correlation; rho = 0.68, $p < 0.05$). Similar correlation between reaction times and activations in brain regions have been demonstrated in auditory perceptual decision-making paradigms (Binder et al., 2004).

Moreover, we examined whether the plateaus simply reflect the amount of time-points over which one needs to average in order to obtain significantly above chance levels results, irrespective of the temporal order of the BF. We therefore randomly permutated the BF within each trial, keeping the same random permutations across trials, but different across the ten test datasets. We then computed the ABF on the permuted trials and performed classification time-point by time-point (Figure 5, red dashed lines). In this case the AUC values were above chance levels on average already at 72ms post-stimulus onset. The difference between classification based on the expanding average of ABF and each permuted version provides striking evidence that the results obtained by keeping the original temporal order within a trial is not a mere consequence of statistical power; accurate discrimination is a result of which points we average over, not how many.

3.5 Time-locked discrimination

When classification was performed at every time-point separately within the trial (Figure 5, green lines), the AUC was never higher that the AUC computed by averaging evidence over time (Figure 5, red solid lines). This was the case for all subjects except

those that had an AUC close to (but above) chance levels (S3 and S6 of the same Figure). This highlights the necessity of considering time-unlocked activity in order to establish an accurate prediction.

3.6 Source estimations

Source estimations were performed on the extracted template maps (one per subject and per experimental condition) and averaged across subjects. Both conditions included prominent sources along the frontal gyrus (BA 11) and middle temporal gyrus (BA 37 for exploitation and BA 21 for exploration; Figure 7a,b). Statistical contrast of these source estimations identified no region more active during exploitation (Figure 7c), further supporting the theory that exploration overrides exploitatory tendencies (Daw et al., 2006). Regions that were significantly more activated during exploration included the right suppramarginal gyrus and the dorsolateral prefrontal cortex (DLPF; BA 9, 46, 47; Figure 7c), known to be involved in decision-making and responsible for task-switching as well as in modulating behavior under uncertainty and deriving conclusions (Hampton and O'doherty 2007; Reverberi et al., 2007; Christopoulos et al., 2009). In particular, the role of the right DLPF in risk-taking (Ernst et al., 2002; Gianotti et al., 2009) and strategic decision making (Knoch et al., 2006) has been argued, here our results suggest that it is involved in risk taking during exploratory decisions.

It is also worth noting that the source estimations were performed on the extracted template maps and not on the actual voltage topographies, as it is usually done in ERP analyses (Michel et al., 2004). As these template maps were extracted from the single-trials irrespective of the time of their appearance it is not possible to assign them to any

specific latency. However, through the Bayes Factor (Figure 2b) we can have an estimate for each trial separately of when in time it is more likely to observe one template map instead of the other.

3.7 Disentangling exploratory decision-making from other confounding factors

To exclude the reward-related confounding factor, we separated the test trials according to wins and losses, but classified them using the exploration/exploitation models. Naturally, wins lead to exploitation and losses to exploration (see also *3.1 Behavior*) and we considered the 'true' labels for this control accordingly. Classification accuracy was on average 0.55 and at chance levels for all but two subjects (Table 1, second row). We remind the reader that the AUC values when classifying the same trials according to exploration-exploitation were on average 0.65 and above chance levels for all but one subject. Moreover, the AUC values when classifying wins/losses were significantly lower than when classifying exploration/exploitation (paired t-test, $p < 0.05$), for four out of seven subjects.

In order to take into account the full range of rewards (which is a continuous variable) and not only binary wins/losses, we also computed the correlation between the discriminant function for every trial, t (DF_t ; Equation 3) and the corresponding rewards (expected – received payoff) and also the corresponding absolute gain for that trial (received payoff). The correlation values were computed for every subject, across all test trials. In the first case, the absolute values of correlation were bellow 0.07 and not significant for any subject (Spearman's $|rho| < 0.07$; $p>0.38$). In the second case, the absolute values of correlation were bellow 0.09 and not significant for six out of seven subjects (Spearman's $|rho| < 0.09$; $p>0.17$). The correlation for the last subject was 0.17 ($p = 0.01$).

The abovementioned results suggest that the GMMs and the extracted template maps are specific to the decision-making and not to reward evaluation. This is consistent with the poor performance of the classifier when extracting the voltage maps at an average level, where the time-periods of differences between exploration/exploitation are likely linked to reward estimation.

On average very few trials corresponded to exploration and also to staying on the same machine (8.7 ± 3 trials per subject during the whole experiment). However, subjects were exploiting while changing machines in a large number of trials (on average 125 ± 31 trials per subject) and we therefore considered only those to eliminate the confound of machine switching. We considered exploitation trials of the validation dataset that forcibly correspond to staying with the same machine and also an equal number of trials that would correspond to exploitation and a change of machine (part of the trials that were labeled as unknown by our behavioral model). Using the extracted template maps for exploration/exploitation we attempted to classify stays as exploitation and switches as exploration. We obtained AUC values above chance levels for 5 out of 7 subjects (paired t-test; $t_{(9)}>4.6$ $p<0.01$) and an average AUC value of 0.59 across subjects (Table 1, third row). However, this high AUC value was mainly driven by the ratio of correctly classified machine stays, which forcibly correspond to exploitation (Table 1, third row, sensitivity = 0.61). The ratio of machine switches that were classified as exploration (specificity) was lower than sensitivity for 5 out of 7 subjects and on average 0.51 (Table 1, third row). This shows that there is possibly an overlap between exploratory decision-making and switching among the machines. We therefore cannot fully dissociate exploration-exploitation from machine switching for all of the subjects, which is also possibly due to power limitations and to the different strategies followed by individual subjects.

4. Discussion

We have shown that it is possible to predict the type of subjects' decision on the next trial using EEG topographic information during reward evaluation. Classification accuracy, measured as the AUC, was above chance levels for six out of seven subjects and on average 0.65. Importantly, our classification results were based on neurophysiologically interpretable features while taking full advantage of the high temporal resolution of EEG. We detected differences between exploratory and exploitative decisions measured as voltage topographies, time-point by time-point. Because different topographies are forcibly a consequence of a change in the configuration of underlying neural networks, we could infer the brain regions that best discriminate between these two types of decisions, while keeping a fine temporal resolution.

For each subject we extracted one representative topography per experimental condition (exploration/exploitation) and performed classification by computing the probabilities over time of appearance of one of them with respect to the other. We observed that time-locked information across trials could not predict the subjects' decisions at any time-point (Figure 5, green lines). An accurate prediction was possible only when taking into account activity over a larger period of time, in accordance to the not time-locked nature of decision-making (Figure 5, red solid lines).

Moreover, we could detect, for every subject separately, the crucial time-period for accurately predicting the subjects' decisions (Figure 5, gray vertical lines). Classification

accuracy did not statistically increase even when considering more activity, after this period. This finding provides a strong indication about the typical time at which the underlying generators of decision-making start differentiating their responses at the single-subject level. In addition, we found that these specific latencies positively correlate with each subject's average reaction times during the experiment (Figure 6 ; rho = 0.68, $p < 0.05$).

In this high level cognitive task different subjects with varying levels of risk-taking can employ different strategies for task completion, and therefore inter-subject variability cannot be neglected. We assumed that the neural correlates underlying exploratory decision-making are similar across subjects, but that they are expressed at different time intervals. For that purpose, we computed and extracted the template maps for each subject separately, but estimated the underlying sources at the group level, revealing the common mechanisms across subjects and allowing us to derive general conclusions (see 4.3 Localization of the relevant generators).

Our results are compatible with what has been shown in similar EEG and fMRI tasks, both in terms of activation of the underlying neural networks (overlapping results with Daw et al., 2006) and of classification accuracy (similar to Hampton and O'Doherty, 2007 based on fMRI and better than Bourdaud et al., 2008 based on EEG, where discriminating was above chance levels for only four out of eight subjects). The main added value of the present work relies on the detection of the relevant time-periods for the decisions; a point to which we will return below.

4.1 Time-locked and unlocked components

The EEG signals were time-locked on the display of reward and thus average ERPs exhibited characteristic responses of prediction error evaluation, peaking at 135ms and 225ms and corresponding to the feedback related negativity and a P300 (Nieuwenhuis et al., 2004; Frank et al., 2005; Yeung et al., 2005; Wu and Zhou 2009). The trials however, are not split into conditions according to the prediction error (i.e. wins/losses), but according to the labels assigned from the behavioral model. Moreover, the analysis of the subjects' behavior revealed that the received reward on trial n-1 indeed influences the subjects' decision on trial n.

At the average ERP level we found three distinct intervals of topographic difference between trials leading to exploration and trials leading to exploitation, similar to what has been previously reported in other decision-making studies (Cohen and Ranganath 2007). However, none of these intervals can lead to an accurate prediction of the subjects' decisions on a trial-by-trial basis (see *3.2 Average ERPs*), indicating that, the neural correlates of single-trial decision-making cannot be found strictly time-locked, across subjects and that inter-trial variability cannot be ignored. We speculate that these time-locked intervals are in fact related to the reward evaluation; especially the two first (~220-250ms and 280-300ms post-stimulus), correspond to the typical latency of characteristic waveforms of such processes (Frank et al., 2005).

So far, other studies have also shown a strong relation between prediction errors and subsequent behavior, either in terms of reaction times, using EEG oscillatory activity (Cavanagh et al., 2010), or in terms of switch/stay decisions, using amplitude differences on average ERPs (Cohen and Ranganath 2007) or single-trial analysis based on linear combinations of electrodes (Philiastides et al., 2010). However, none of them reports

quantitative results on predicting the subjects' decisions, nor fully analyze the temporal aspects of discrimination between the subjects' decisions.

4.2 Relevant time-periods for decision-making

The typical time latencies at which we could already accurately predict the subjects' decisions were on average at 510ms after the display of the reward. This does not mean, however, that decisions are made exactly at 510ms, but rather that they are already made at some point before. A stricter statistical threshold could have possibly detected an even earlier point, but this would still be consistent with our results. The periods detected here in fact corresponded to plateaus in the AUC values (see *3.4 Relevant time-periods for decision-making*) and appeared much later than sensory-evoked processing of the displayed payoff, and the reward-related activity (feedback-related negativity), which typically peaks around 150-300ms after feedback presentation (Holroyd and Coles, 2002). Although these plateaus correlated significantly with the subjects' average reaction times, they occurred on average at ~880ms before the subsequent button presses. Moreover, due to the task design, subjects were forced to wait for 1s after the display of reward so it is unlikely that the detected plateaus are trivially related to motor responses.

In a previous study using the same subjects and paradigm (Bourdaud et al., 2008) it was shown that it is also possible to find discriminating oscillatory patterns of EEG activity of the current decision within 1000ms prior to button press. Consistently, we found an overlap between the time period that was used there with the time period that is used in the present study. However, our results provide finer temporal information, estimated at the single-subject level. Importantly, we showed that it is possible to discriminate between the two types of decisions already from ~510ms after the presentation of the

reward in the previous trial. We therefore exclude that all the crucial information for establishing that prediction is found right before the button-press. This finding provides a strong indication for future studies for defining an optimal trial duration when investigating decision-making.

More generally, assessing the time when decisions are made has been the subject of previous studies, by means of fMRI (Soon et al., 2008), or pupil dilation (Einhäuserat al., 2010). In perceptual decision-making it has been shown that the higher the degree of uncertainty, the longer it takes for subjects to reach a decision (Heekeren et al., 2004), similar to what is now known for cost-benefit-based decision making (Basten et al., 2010). As a future direction, we can directly manipulate decision times by employing a similar paradigm in reward-based decision making and observe the the effects of such a manipulation on the EEG responses. Moreover, fMRI activations in relevant regions of interest in perceptual decision-making are known to correlate with the subjects' response times (Binder et al., 2004; McKeeff and Tong, 2006) and this correlation is used as a proof for the relevance of the regions of interest with decision-making processes. In a similar way, we demonstrate here a correlation between temporal features of the EEG and subjects' reaction times, to further support the validity of our results.

The advantage of the present approach is that we can examine the variability across subjects, of the relevant time-periods for the decisions, even under identical experimental conditions. To the best of our knowledge this is the first EEG study to determine with minimal *a priori* temporal constraints, at the millisecond time-scale, the single-subject relevant time-periods for differentiating neural activity for the following decisions.

4.3 Localization of the relevant generators

Source estimations were performed on the extracted template maps and averaged across subjects. No region was more active during exploitation when statistically contrasting the two conditions. However, the right suppramarginal gyrus and the right DLPF were significantly more active during exploration. The DLPF has been previously linked with task-switching, decision-making and behavior under uncertainty (Hampton and O'doherty 2007; Reverberi et al., 2007; Christopoulos et al., 2009; Gianotti et al., 2009).

However, in a similar fMRI study (Daw et al., 2006), it has been shown that the frontopolar cortex and the intraparietal sulcus were more active during exploration. The frontopolar cortex has also been reported to be responsible of evidence integration and to be functionally connected to parietal and premotor regions, during a similar decision making task (Boorman et al., 2009). A possible reason for this discrepancy between our study and fMRI results is that our sources are estimated relatively early in time and under a substantially different time-scale. We speculate that the regions reported in Daw et al. and namely the frontopolar cortex, are responsible for gathering activations from the DLPF and the supramarginal gyrus, (the regions we obtain here) in order to eventually lead to exploratory decisions.

4.4 Disentangling decision-making from other confounds

As expected, when subjects experience a loss they are more likely to explore on the next trial and when they experience a win they are more likely to exploit. However, we demonstrated that our results are not only reflecting the discrimination between wins

and losses, as we were not able to accurately classify wins versus losses using the trained models for exploration/exploitation for most of the subjects (5/7).

Our behavioral model was restricted so that exploration always corresponds to switching machine and exploitation to staying with the same machine, in order to obtain directly comparable results with the previous study using the same paradigm and subjects (Bourdaud et al., 2008). Machine switching and exploration/exploitation appear to be strongly intermixed as it was also clear from the subjects' behavior, as exploration was combined with a machine stay in less than nine trials per subject during the whole experiment, for most of the subjects.

Conclusions

In summary, we show that by using EEG topographic activity during reward evaluation it is possible to accurately predict subsequent decisions. The neural correlates of the subjects' decisions can be detected as early as ~880ms before the button press and are localized within the suppramarginal gyrus and the right DLPF.

References

Basten U, Biele G, Heekeren HR, Fiebach CJ. (2010). How the brain integrates costs and benefits during decision making. Proc Natl Acad Sci U S A.

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD. (2004). Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci. 7(3):295-301.

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron. 62(5):733-43.

Bourdaud N, Chavarriaga R, Galan F, Millan Jdel R. (2008). Characterizing the EEG correlates of exploratory behavior. IEEE Trans Neural Syst Rehabil Eng. 16(6):549-56.

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PG, Kwok SC, Phillips A, Tanaka K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science. 325(5936):52-8.

Cavanagh JF, Frank MJ, Klein TJ, Allen JJ. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. Neuroimage. 49(4):3198-209.

Christopoulos GI, Tobler PN, Bossaerts P, Dolan RJ, Schultz W. 2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. J Neurosci.29(40):12574-83.

Cohen JD, McClure S.M, Yu AJ (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Phil. Trans. R. Soc. B. 362(1481):933-942.

Cohen MX, Ranganath C. (2007). Reinforcement learning signals predict future decisions. J Neurosci. 10;27(2):371-8.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. (2006). Cortical substrates for exploratory decisions in humans. Nature 441(7095):876-9.

De Lucia, M., Michel, C.M., Clarke, S., and Murray, M.M., (2007). Single-trial topographic analysis of human EEG: A new 'image' of event-related potentials. Proceedings Information Technology Applications in Biomedicine.

Einhäuser W, Koch C, Carter OL. (2010). Pupil dilation betrays the timing of decisions. Front Hum Neurosci. 26;4:18.

Ernst M, Bolla K, Mouratidis M, Contoreggi C, Matochik JA, Kurian V, Cadet JL, Kimes AS, London ED. (2002). Decision-making in a risk-taking task: a PET study. Neuropsychopharmacology. 26(5):682-91.

Frank MJ, Woroch BS, Curran T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. Neuron.47(4):495-501.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 104(41):16311-6.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F., (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci. 12(8):1062-8.

Gianotti LRR, Knoch D, Faber PL, Lehmann D, Pascual-Marqui RD, Diezi C, Schoch C, Eisenegger C, Fehr E, (2009). Tonic activity level in the right prefrontal cortex predicts individuals' risk taking. Psychol Sci 20:33–38.

Gold JI, Shadlen MN., (2007). The neural basis of decision making. Annu Rev Neurosci. 30:535-74.

Green D.M., Swets J.M., (1966). Signal detection theory and psychophysics. New York: John Wiley and Sons Inc.

Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG. (2004). A general mechanism for perceptual decision-making in the human brain. Nature. 431(7010):859-62.

Hampton AN, O'doherty JP. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. Proc Natl Acad Sci U S A. 104(4):1377-82.

Holroyd CB, Coles MG. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev.109(4):679-709.

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science. 310(5754):1680-3.

Jepma M. and Nieuwenhuis S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. J Cogn Neurosci. 23(7):1587-96.

Kaelbling LP, Littman ML, Moore AW. (1996). Reinforcement learning: a survey. J. Artif. Intel.Res. 4:237–85

Knoch D, Gianotti LR, Pascual-Leone A, Treyer V, Regard M, Hohmann M, Brugger P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. J Neurosci. 26(24):6469-72.

Koenig T, Melie-García L., (2010). A method to determine the presence of averaged event-related fields using randomization tests. Brain Topogr. 23(3):233-42.

Lehmann D, Skrandies W (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalogr Clin Neurophysiol 48:609–621.

McKeeff TJ, and Tong F. (2006). The Timing of Perceptual Decisions for Ambiguous Face Stimuli in the Human Ventral Visual Cortex. Cerebr. Cortex., 17(3) : 669-678.

Michel CM, Murray MM, Lantz G, Gonzalez S, Spinelli L, Grave de Peralta R., (2004). EEG source imaging. Clin Neurophysiol. 115(10):2195-222.

Murray MM, De Lucia M, Brunet D, Michel CM (2009) Principles of Topographic Analyses of Electrical Neuroimaging. In: Event-Related Potentials II: Advances in ERP, EEG, & MEG Analysis MIT Press (Handy TC, Ed).

Nieuwenhuis S, Holroyd CB, Mol N, Coles MG. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. Neurosci Biobehav Rev. 28(4):441-8.

Pascual-Marqui RD, (1999). Review of methods for solving the EEG inverse problem, Int J Bioelectromagn 1 (1), pp. 75–86.

Pascual-Marqui RD, Michel CM, Lehmann D, (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain, Int J Psychophysiol 18 (1) pp. 49–65.

Penny WD, Stephan KE, Mechelli A, Friston KJ, (2004). Comparing dynamic causal models. Neuroimage 22(3):1157-72.

Perrin F, Pernier J, Bertrand O, Giard MH, Echallier JF, (1987). Mapping of scalp potentials by surface spline interpolation. Electroencephalogr Clin Neurophysiol 66:75–81.

Philiastides MG, Biele G, Vavatzanidis N, Kazzer P, Heekeren HR. (2010). Temporal dynamics of prediction error processing during reward-based decision making. Neuroimage 53(1):221-32.

Platt ML, Glimcher PW. (1999). Neural correlates of decision variables in parietal cortex. Nature. 400(6741):233-8.

Raftery A.E., (1995) Bayesian Model Selection in Social Research. Sociological Methodology, 25: 111-163.

Ratcliff R, Philiastides MG, Sajda P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proc Natl Acad Sci U S A. 106(16):6539-44.

Reverberi C, Cherubini P, Rapisarda A, Rigamonti E, Caltagirone C, Frackowiak RS, Macaluso E, Paulesu E.(2007). Neural basis of generation of conclusions in elementary deduction. Neuroimage 38(4):752-62.

Rushworth MF., Behrens TEJ. (2008b) Choice, uncertainty and value in prefrontal and cingulate cortex. Nature Neurosc. 11, 389 – 397.

Seo H, Barraclough DJ, Lee D. (2009). Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J Neurosci. 29(22):7278-89.

Soon CS, Brass M, Heinze HJ, Haynes JD. (2008). Unconscious determinants of free decisions in the human brain. Nat Neurosci. 11(5):543-5.

Sul JH, Kim H, Huh N, Lee D, Jung MW. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron. 66(3):449-60.

Suton R., Barto A., (1998). Reinforcement learning: an introduction. Cambridge, MA: MIT Press.

Talairach J, Tournoux P., (1988). Co-planar stereotaxic atlas of the human brain. New York: Thieme.

Tzovara A, Murray MM, Plomp G, Herzog MH, Michel CM, De Lucia M (2011). Decoding stimulus-related information from single-trial EEG responses based on voltage topographies. *Pattern Recognition*. doi:10.1016/j.patcog.2011.04.007

van 't Wout M, Kahn RS, Sanfey AG, Aleman A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. Neuroreport. 16(16):1849-52.

Wong KF, Wang XJ. (2006). A recurrent network mechanism of time integration in perceptual decisions. J Neurosci. 26(4):1314-28.

Wu Y, Zhou X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. Brain Res. 1286:114-22.

Yeung N, Holroyd CB, Cohen JD. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. Cereb Cortex. 15(5):535-44.

Yoshida W, Ishii S. (2006). Resolution of uncertainty in prefrontal cortex. Neuron. 50(5):781-9.

Figure Captions

Figure 1

(a) Experimental protocol. Each trial is comprised of three phases. First the four machines are presented and participants have 1s to select one of them by pressing a key (choice phase). Once a machine has been chosen the rest of them are deactivated (grayed) for 1s (delay phase). Finally the payoff for the selected machine is displayed for 1s (display phase).

The red thick line shows how we define a trial for the current study: trials were extracted with respect to the display of reward, spanning 100 ms before the display and 780 ms post-stimulus onset. Activity from this period was used to predict the subjects' decision on the next trial.

(b) Example of the payoff evolution across the experiment for the four machines.

Figure 2

Template maps and their Bayes Factor across trials for an exemplar subject. (a) Template maps for exploitation (left panel) and exploration (right panel), averaged across shuffles. (b) Bayes Factor for each trial. The Bayes Factor is computed for each of the topographies as the ratio of the posterior probability of the template map for exploitation divided by the posterior probability of the template map for exploration (see Equation 2).

Figure 3

Average ERP waveforms across subjects of electrode Cz and characteristic scalp topographies at 135 and 225 msec for the two conditions (blue line exploitation, green line exploration). The periods of topographic difference between the two conditions identified by the TANOVA are highlighted in red on the x-axis. Only periods identified in at least 8 out of 10 splits of the data are displayed here.

Figure 4

Classification performance for each subject based on the overall trial (light gray bars), and chance levels (dark gray bars). The mean AUC value across subjects was 0.65. Chance levels were determined by randomly shuffling 100 times the true labels of the test trials and then classifying them. On average, chance AUC values were 0.5. Asterisks show the subjects whose AUC values were significantly above chance levels (all but S5).

Figure 5

Time-course of AUC values based on the expanding average version of the BF. In solid red we show the temporal evolution of the AUC values if, for every time-point, we take into account activity from the beginning of the trial up to that time-point. Acuracy increases in time as we add more evidence until a plateau is reached, then it no longer changes significantly (gray vertical lines). The gray lines indicate at what point in time the AUC values do no longer increase significantly. The red dashed lines correspond to the AUV values computed over a permuted version of the BF. In that case the plateau is reached much earlier (~72ms on average). AUC values are at chance levels when computed time-point by time-point (green lines).

Figure 6

Average reaction times for every subject versus the onset of the detected plateaus in AUC values. Even if we consider more activity after the onset of these plateaus the AUC values do not increase significantly, indicating that the decisions have already been made. These points significantly correlate with the subjects' reaction times, averaged across the experiment (rho = 0.68, $p < 0.05$).

Figure 7

Source estimations on the template maps across subjects for exploitation (a) and exploration (b). Results are rendered on the average MNI brain. Axial slice shows the activations for each of the two conditions in correspondence to the maximal t value at 48, -50, 27 mm. (c) Results of the statistical contrast of the source estimations between exploration and exploitation (paired t-test, $p < 0.05$).

Table 1

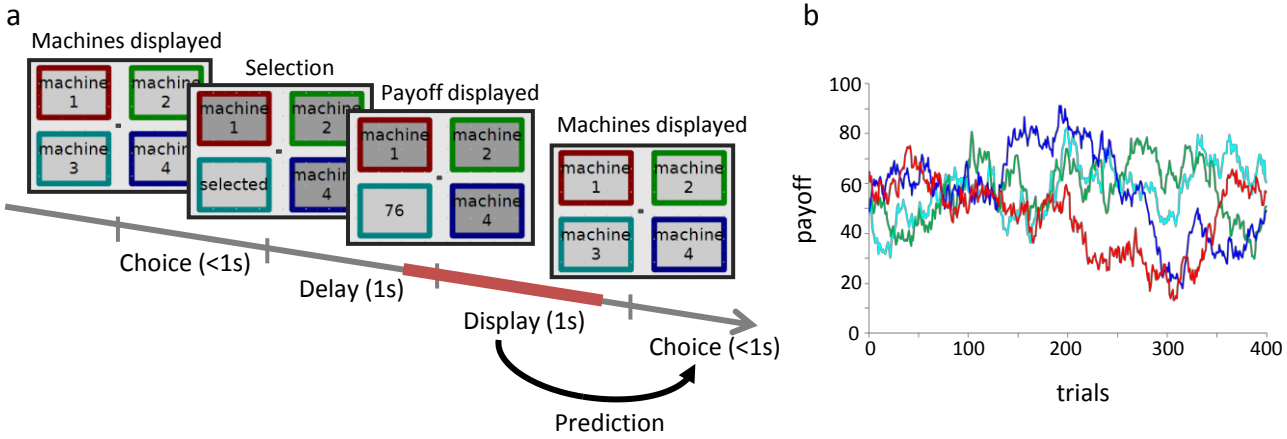
Summary of the classification results, on average across subjects. The columns display AUC values, sensitivity, specificity, the number of trials per condition that entered the comparison and the number of subjects for which we obtained above chance levels results. The first line corresponds to discriminating between exploration and exploitation. The second line shows classification results when keeping the GMMs from exploration/exploitation but splitting the test trials according to wins/losses. In the third line we only consider exploitatory validation trials and keeping the GMMs from exploration/exploitation we classify switch/stays.

8. Table1

	AUC	Sensitivity	Specificity	# trials	# above chance
Exploration Exploitation	0.65	0.63	0.56	(125/125)	6/7
Wins Loses	0.55	0.54	0.55	(133/109)	2/7
Switch Stay	0.59	0.61	0.51	(40/40)	5/7

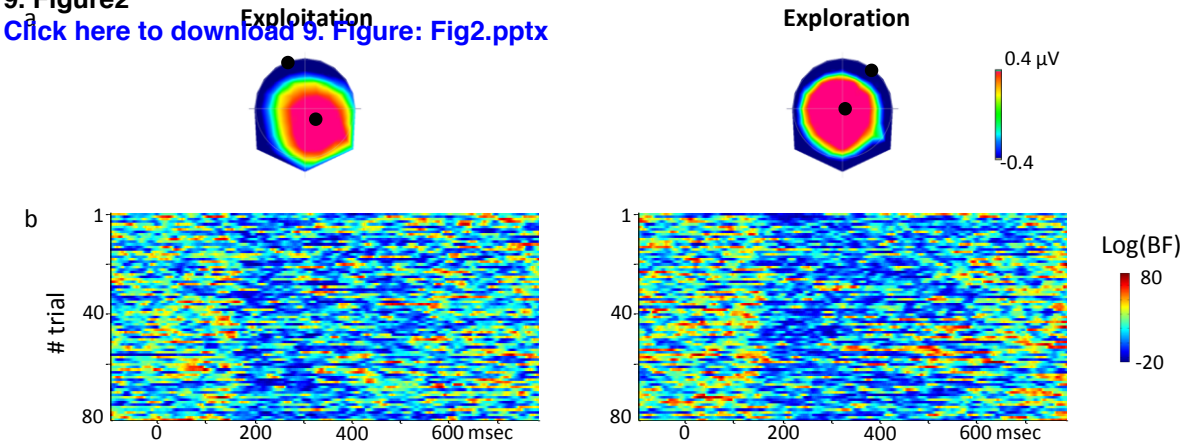
9. Figure1

Click here to download 9. Figure: Fig1.pptx



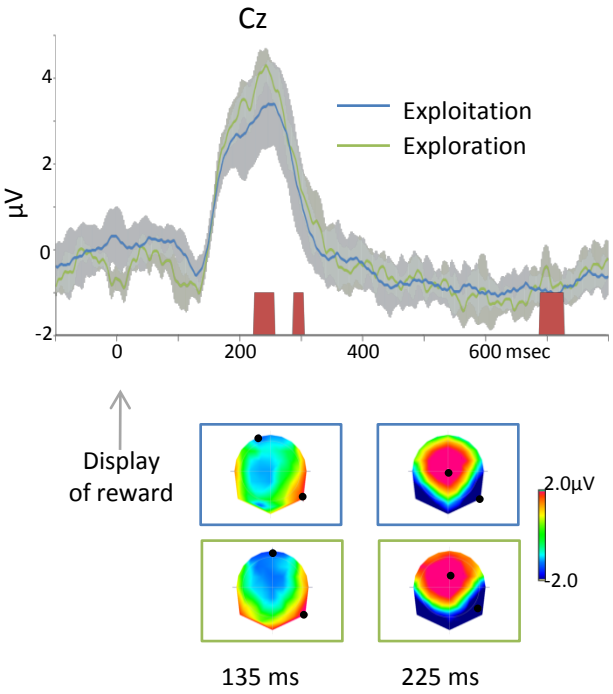
9. Figure2

Click here to download 9. Figure: Fig2.pptx



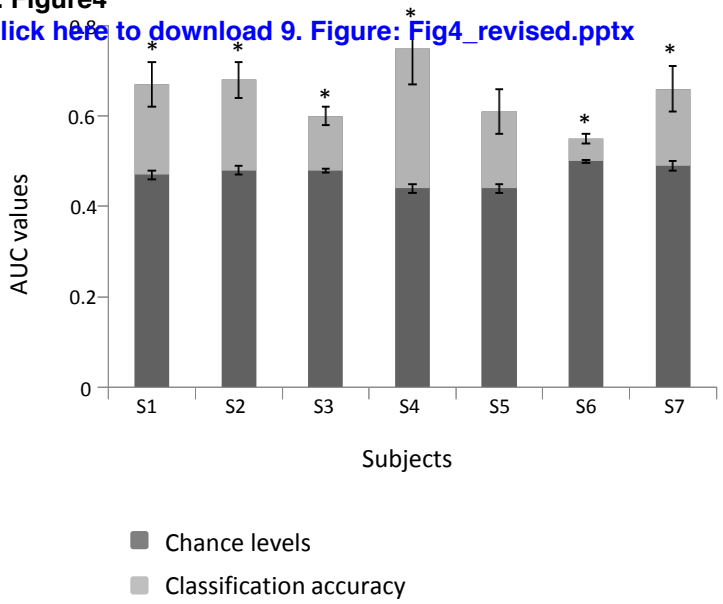
9. Figure3

Click here to download 9. Figure: Fig3_revised.pptx



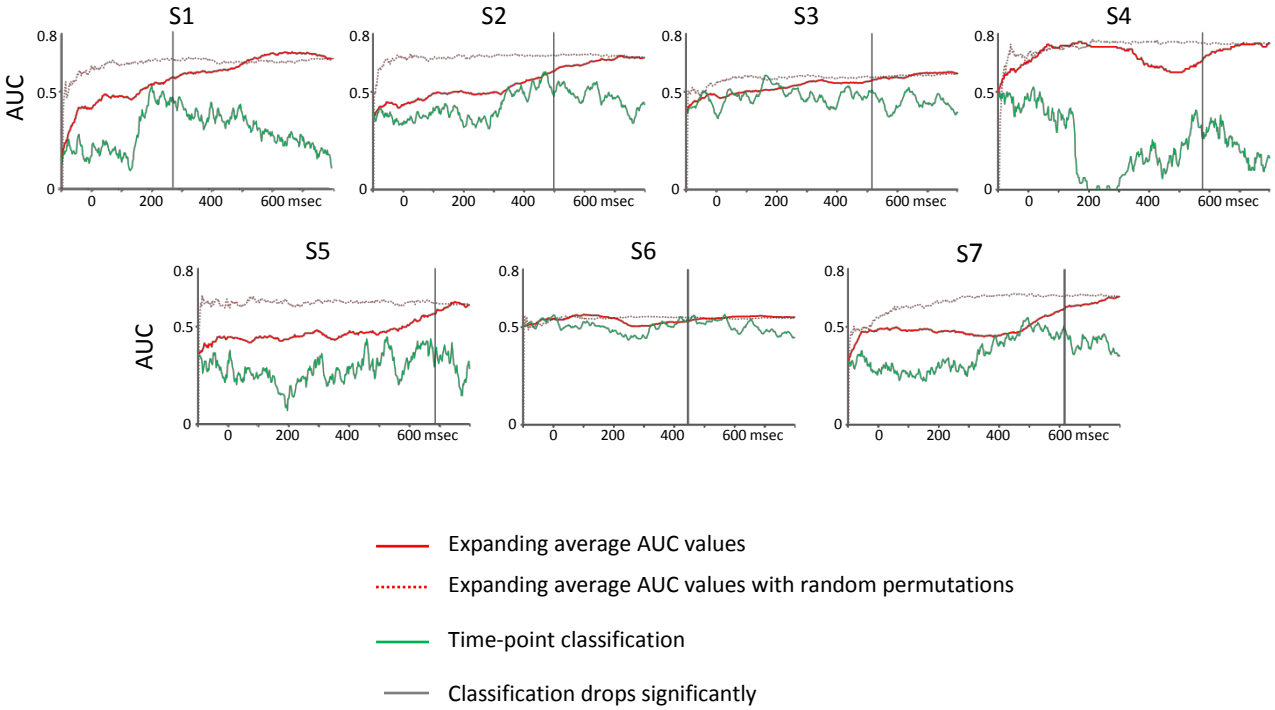
9. Figure4

Click here to download 9. Figure: Fig4_revised.pptx



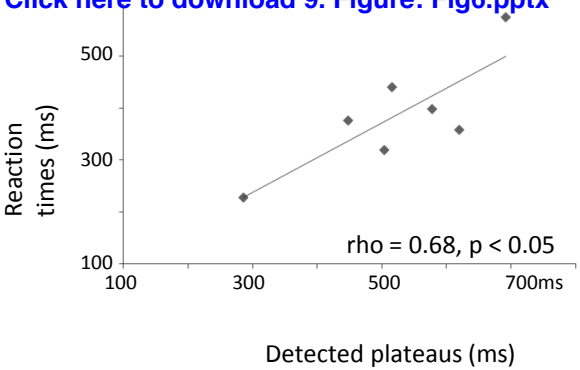
9. Figure5

Click here to download 9. Figure: Fig5.pptx



9. Figure6

Click here to download 9. Figure: Fig6.pptx



9. Figure7

Click here to download 9. Figure: Fig7.pptx