Pallidal neuromodulation of the explore/exploit trade-off in decision-making

Ana Luísa de A. Marcelino[1,2], Owen Gray[3], Bassam Al-Fatly[1], William Gilmour[3], J Douglas Steele[3], Andrea A. Kühn[1,2,4,5,6], Tom Gilbertson[3,7*]

1 Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Movement Disorder and Neuromodulation Unit, Department of Neurology, Charité Campus Mitte, 10117 Berlin, Germany

2 Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Facility Genomics, Charitéplatz 1, 10117 Berlin, Germany

3 Division of Imaging Science and Technology, Medical School, University of Dundee, Dundee, UK

4 Berlin School of Mind and Brain, Charité - University Medicine Berlin, Berlin, Germany

5 NeuroCure, Charité - University Medicine Berlin, Berlin, Germany

6 DZNE, German Centre for Degenerative Diseases, Berlin, 10117, Germany

7 Department of Neurology, Ninewells Hospital & Medical School, Dundee, UK

*Corresponding Author:-

Tom Gilbertson
Department of Neurology
Ninewells Hospital & Medical School, Dundee, UK, DD1 9SY
t.gilbertson@dundee.ac.uk

Word count: 6358

Running Title: Effects of GPI-DBS on decision making

Keywords: Basal Ganglia, Deep Brain Stimulation, Neuromodulation, Globus Pallidus Interna, Explore-Exploit dilemma

Abstract

Every decision that we make involves a conflict between exploiting our current knowledge of an action's value or exploring alternative courses of action that might lead to a better, or worse outcome. The sub-cortical nuclei that make up the basal ganglia have been proposed as a neural circuit that may contribute to resolving this explore-exploit 'dilemma'. To test this hypothesis, we examined the effects of neuromodulating the basal ganglia's output nucleus, the globus pallidus interna, in patients who had undergone deep brain stimulation (DBS) for isolated dystonia. Neuromodulation enhanced the number of exploratory choices to the lower value option in a 2-armed bandit probabilistic reversal-learning task. Enhanced exploration was explained by a reduction in the rate of evidence accumulation (drift rate) in a reinforcement learning drift diffusion model. We estimated the functional connectivity profile between the stimulating DBS electrode and the rest of the brain using a normative functional connectome derived from heathy controls. Variation in the extent of neuromodulation induced exploration between patients was associated with functional connectivity from the stimulation electrode site to a distributed brain functional network. We conclude that the basal ganglia's output nucleus, the globus pallidus interna, can adaptively modify decision choice when faced with the dilemma to explore or exploit.

Introduction

Flexible adaption of choice behaviour is crucial for decision-making when facing uncertainty. The balance between exploiting familiar options and exploring alternatives with potentially less favorable outcomes, represents a fundamental challenge of adaptive control [1]. The neural basis of how humans resolve this dilemma has been advanced significantly over the last decade [2, 3], but the specific contribution from sub-cortical nuclei that make up the basal ganglia (BG) remains poorly defined [4, 5].

To investigate how humans solve this conflict in reinforcement learning, so-called 'bandit tasks' have been used, where subjects choose between several options associated with different reward probabilities to maximize reward payoff. In a 'reversal-learning' paradigm, the reward probabilities change (i.e. reverse), thus subjects need to constantly evaluate whether choice feedback represents a reversal false alarm (i.e. a lower than expected outcome from the option with the highest reward probability) or a true reversal in the reward probabilities. Even in reversal-learning tasks with low levels of payout uncertainty, human participants make a significant number of choices to the lower value option, rather than the choice with the higher expected value. This leads to the observation of probability matching [6], where humans fail to maximise their payoff and adopt a sub-optimal policy which matches their response probabilities to that of the payoff probability. 'Non-greedy' choices, are by definition exploratory, as they are made to the option with the lower expected value [7]. Conversely, 'greedy' choices are made to the option with the highest value [8-10] and can be considered to exploit current knowledge of the values of the actions [11]. One possibility is that exploration represents a cognitive strategy for active information seeking, to reaffirm existing beliefs in the relative value of the choices [3, 10, 11]. Alternatively, it may represent random choice perturbations caused by limitations in the brain's implementation of the decision process leading to decision noise [12]. Either way, the strategies adopted by human participants performing bandit tasks, including reversal-learning [9, 13, 14], rely on exploiting stability whilst at the same time adapting to volatility by exploring alternative courses of action.

Several observations support a role of the BG circuit in explore-exploit decisions. Computational circuit models of the BG predict that its principal function is action selection - the process where by one action (or decision) is selected over other competing alternatives [15, 16]. An extension of this action selection function is the proposal that it's precision (i.e., how selective the circuit is) can be influenced by the overall excitability of the BG circuit [17]. In the context of the explore-exploit dilemma, this means the BG may implement a neural 'decision filter,' with the bandwidth of this filter being adjusted by neuromodulators, such as dopamine, which influence BG excitability [18]. The same models also predict that the excitability of the basal ganglia's principle output nucleus, the globus pallidus interna (GPI) should be a read-out of whether the BG are supporting an exploratory or exploitative decision strategy [19, 20]. Consistent with this prediction, the firing rate of GPI neurons in non-human primates encodes the transition from exploratory to exploitative decision making [5]. Transient decreases in GPI firing observed during learning have been proposed as a potential source of decision variability and

could therefore represent a neurophysiological correlate of exploratory (non-greedy) choices observed in bandit tasks.

At a cortical level, functional imaging (fMRI) supports a role for distinct pre-frontal cortical regions in reversal-learning including the orbito-frontal cortex [21-24], ventromedial prefrontal (vmPFC), dorsolateral prefrontal (DLPFC) and anterior cingulate (ACC) cortices. The same pre-frontal regions are also implicated in explore-exploit choices using (so-called 'restless') bandit tasks with high between-trial outcome volatility and associated choice uncertainty [2, 7]. Analysis of brain activation patterns in these tasks support distinct cortical regions are involved in the different types of decisions. Consistent with their role within a valuation network, ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC) [2, 7, 12, 25, 26] are activated during exploitative choices. In contrast, exploratory choices activate frontopolar cortex (FPC), dorsolateral prefrontal cortex (DLPFC), anterior insula and anterior cingulate cortex, commensurate with their proposed functions in the encoding of uncertainty, behavioural switching and cognitive control [27-31]. The striatum, as the BG input nucleus, receives afferents from these same pre-frontal cortical regions forming segregated cortico-striatal-thalamo-cortical circuits which respect the functional anatomy of distinct cortical circuits for explore-exploit decision making, thus serving as a point of convergence [32].

To date, no study has directly addressed the hypothesised role of the GPI in the human approaches to resolving explore-exploit decisions during reversal-learning. Deep Brain Stimulation (DBS) of the GPI is routinely performed as a treatment for isolated generalised dystonia and also in patients with focal and segmental dystonia if botulinum toxin treatment fails [33]. Dystonia is a movement disorder characterized by sustained or intermittent muscle contractions causing abnormal postures. In focal dystonia it is restricted toone region such as the head or neck [34]. Patients with dystonia have normal levels of striatal dopamine, intact reward prediction error (RPE) signalling [35], and exhibit subtle cognitive biases [36]. These include abnormalities of reinforcement learning (RL) such as increased risk taking and delayed flexibility to changes in reward contingency reversal [35, 37]. With the caveat of these disease specific RL abnormalities in mind, dystonia patients with chronically implanted DBS-GPI electrodes represent a unique opportunity to test the influence of the GPI and, in turn, as its principle output nucleus, the basal ganglia's role in explore-exploit decision making in humans.

Our working hypothesis was that neuromodulation of the excitability of GPI neurons would modify explore-exploit decisions in a reversal-learning task. As DBS suppresses firing rates in human GPI neurons [38, 39] we hypothesised that DBS mediated reduced GPI excitability should drive greater proportion of exploratory choices consistent with previous findings by Sheth et. al., 2011[5]. To test this hypothesis, we examined choice behaviour in patients with chronically implanted DBS electrodes in both "ON" and "OFF" stimulation whilst they executed a 2-armed bandit reversal-learning task. Given the more recognised role of the BG in habitual learning of stimulus-response associations [40, 41], we purposely chose a low volatility probabilistic reversal-learning task to engage cortico-sub-cortical circuits in "model-free" learning [42]. Our choice of this task was also aimed at isolating the specific contribution of the GPI to random exploration. This form of exploration is more closely aligned with the

sub-optimal, non-greedy choices identified during pauses in GPI firing in primates [5]. Exploration can also be “directed” to the option which strategically aims to minimise uncertainty about a choice that a decision maker has least familiarity with [2]. Accordingly, directed exploration relies upon brain regions that encode working memory [28] and circuits including the external segment of the globus pallidus (GPe) which contribute to active information seeking to resolve uncertainty [43],

By fitting a model of the decision-making process (Reinforcement Learning Drift Diffusion Model: RLDDM) we aimed to test further mechanistic hypotheses regarding latent cognitive effects of DBS neuromodulation on GPI choice arbitration [44]. The RLDDM incorporates both decision choices and decision time (DT) information, thus affording a richer interpretation of the mechanisms underlying choice compared to more traditional reinforcement learning models

We further hypothesised that the influence of GPI-DBS neuromodulation in decision-making should be dependent on the connectivity of the stimulated site with cortical areas implicated in explore–exploit choice behaviour.

Results

Eighteen patients with isolated dystonia (for clinical details see Supplementary Table 1) who had chronically implanted DBS electrodes (Figure 1A) targeted at the globus pallidus interna (GPI) performed a reversal-learning task (Figure 1B&C). The patients were instructed to try and win as many “vouchers” as possible throughout the duration of the task by choosing one of two options presented to them on a computer screen. Patients were tested with their DBS device turned “ON” or “OFF” in a randomised order between the two task blocks. Each block consisted of three, forty-trial sessions where the probability associated with winning a “voucher” switched from 80%:20% to 20%:80% midway through the second session (Figure 1B). Participants performed two versions of this task separated by an interval of 20min after the initially randomized DBS condition (ON or OFF) had been inverted. The patient’s performance was compared with a group of 18 age- and sex-matched healthy controls (HC) performing task block 1. The two blocks differed only by the fractals presented (Figure 1B). Participants were not informed about the contingency reversal in the task.

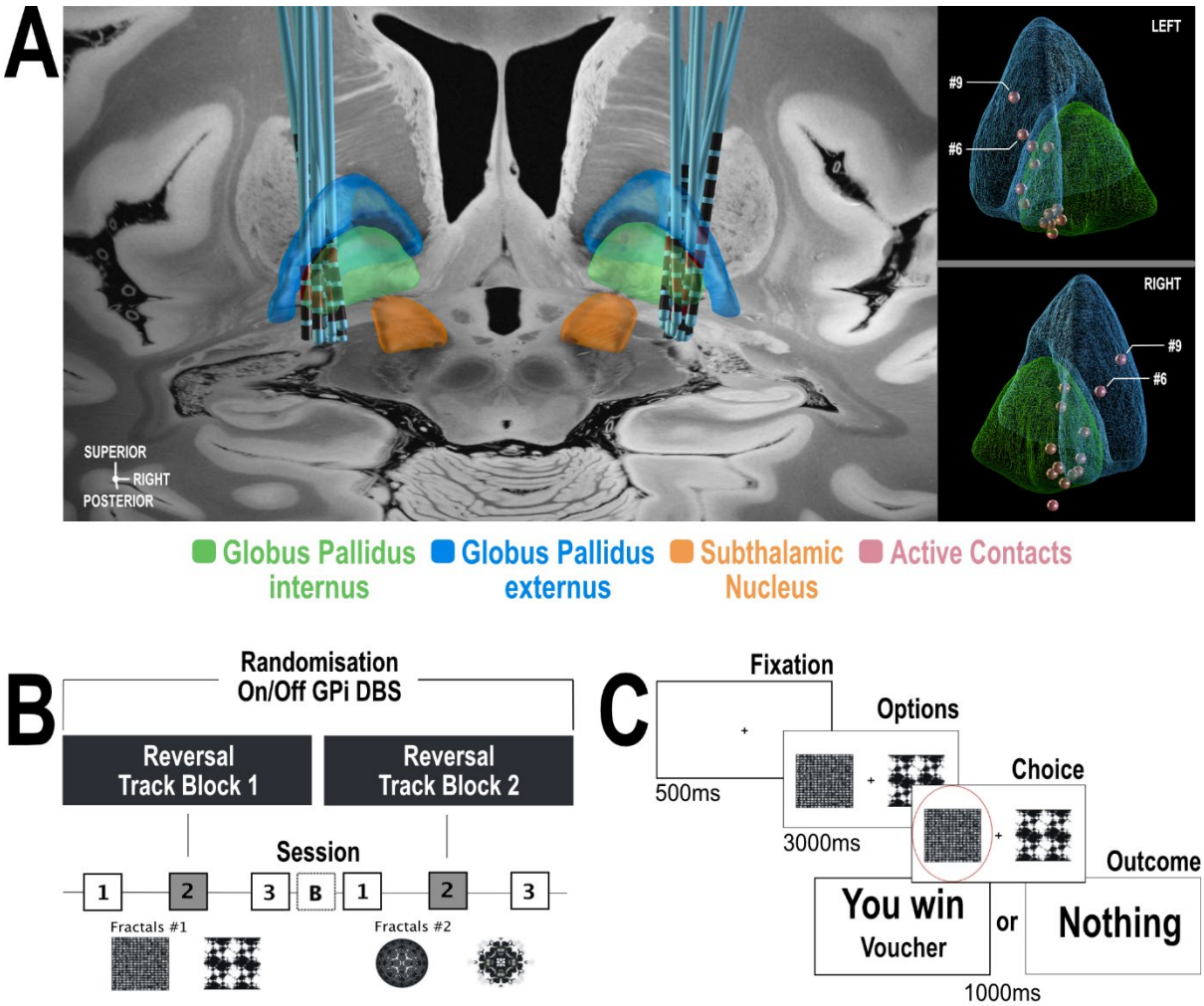

Figure 1: General overview of the study design and task. (A) Localization of DBS electrodes was done with Lead-DBS (www.lead-dbs.org) as previously described [46]. The left panel depicts electrode localization within the GPI projected to a 7T brain backdrop[45]. On the right panel, active contacts location is shown in red – only two patients (# 6 and #9) show active contacts within the GPe (blue). **(B)** Each of the two reversal learning task blocks. The blocks differed only in the pair of fractals used as stimulus options. Each reversal learning task block consisted of three 40 trial sessions. Reversal of the probability of receiving a reward occurred half-way through session two. The task was performed once in the OFF- DBS state and once in the ON-DBS state in a counterbalanced manner with a 20-minute break ('B') in between blocks 1 and 2. During this, DBS stimulator was either switched ON or OFF. **(C)** Example of a single trial in the reversal learning task. On each trial, subjects chose either left or right fractal options, which were also counterbalanced, using their left or right hand to press the corresponding keyboard button. The selected cue was then shown surrounded by a red circle (in this example Task Block 1 the left-hand cue is chosen). Subjects were then presented with the outcome of their choice on the next screen, which could be either a reward ('You Win') or zero ('Nothing'). Outcome probabilities of receiving a reward on choosing either fractal were 80%:20%.

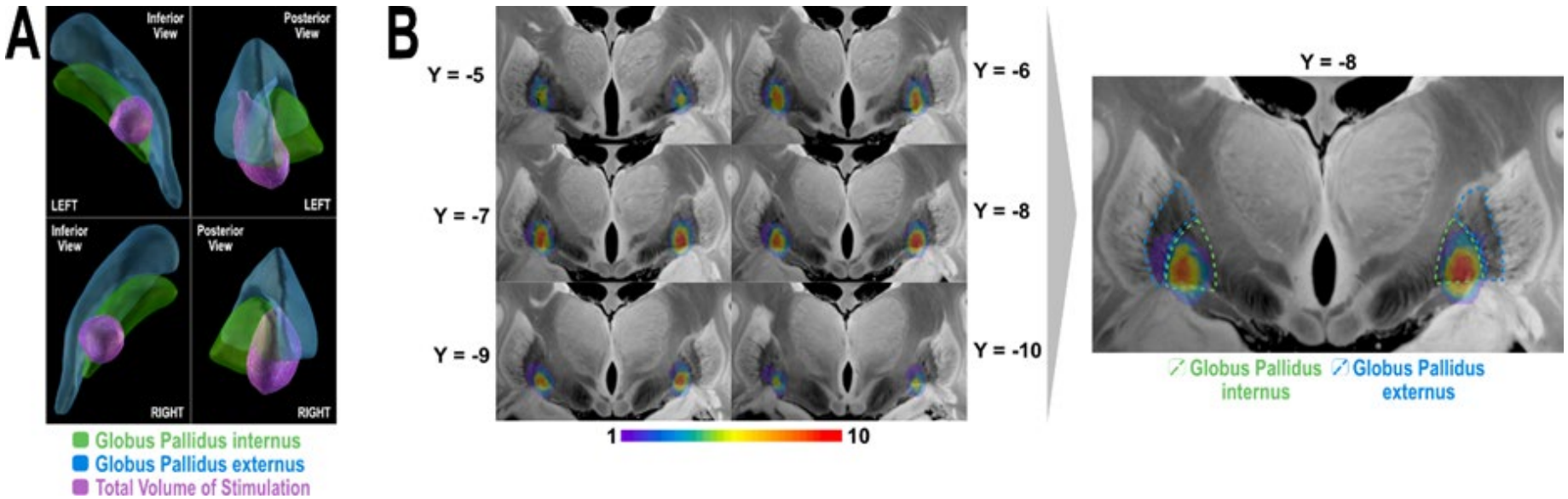

Figure 1 - Figure Supplement 1: Spatial distribution of stimulation volume on a group level. (A) 3D rendering of total volume of stimulation (purple) summed for n = 14 patients illustrated in relation to the GPi (green) and GPe (blue) nuclei. (**B)** 2D coronal slices [45] of n-map representing voxel-wise heat-map of how many stimulation volumes are overlapping in each voxel of the total stimulation volume. The slices are selected to cover the extent of the total stimulation volume from most anterior to most posterior aspect. Of note, the majority of the stimulation volumes were located in the posteroventral part of the GPi as indicated by the red-yellow color gradient of the heat-map. Voxels overlapped with GPe were covered by ~1-2 stimulation volumes (purple-blue color) and mainly encroaching on its medial margin (right panel of B). Weighted-sum of overlap between the heat-map and the GPI is 1556.5, while that with the GPe is 376.5. The weighted-sum GPi/GPe ratio of overlap is 4.13, which means that GPi is being stimulated ~4 times more frequently than GPe in our cohort. The maximal overlap value in the heat-map is 10 which is less than patients' number (n = 14). This is because of the fact that the heat-map distribution is governed by spatial heterogeneity of the stimulation volumes and their geometrical configuration.

GPI DBS enhances exploratory choices without affecting task performance

A fixed effects analysis of performance (number of vouchers won) was conducted, using a two-way repeated measures ANOVA with stimulation state (ON-DBS versus OFF-DBS) and task session as fixed factors. Consistent with the detrimental effect of contingency reversal on performance in session 2, there was a main effect of session on performance ($F(2,34) = 6.26$, $p = 0.002$, Average rewards: Sessions 1, 2 and 3 ; 27.5±0.8, 23.6±0.6, 24.8±0.9) but no main effects of DBS state ($F(1,34) = 0.9$, $p = 0.34$) or interaction between DBS state and performance in a session of the task ($F(2,34) = 1.03$, $p = 0.30$) see Figure 3 A&B. The same analysis applied to the decision time (defined as the interval between the choices being presented and keypress) confirmed a significant main effect of session ($F(2,34) = 5.21$, $p = 0.005$, DTs: Sessions 1, 2 and 3 ; 1.04±0.04, 1.14±0.05, 1.18±0.05 seconds) but no effect of DBS ($F(1,34) = 0.01$, $p = 0.92$) or interaction between DBS and the decision times within a session of the task ($F(2,34) = 1.27$, $p = 0.27$). This analysis confirmed that switching their DBS stimulator to the OFF state did not affect task performance by deteriorating the patient's symptom control. Appling the same analysis, with group (HC or patients) and task session as fixed factors, the number of rewards obtained in each session by the patients was no different compared with the HC group (HC average rewards: Sessions 1, 2 and 3 ; 24.7±1.0, 22.7±0.8,28.1±0.8; $F(1,70) = 0.01$, $p = 0.91$) but the decision times in the HC group were on average faster compared to patients (HC DTs: Sessions 1, 2 and 3 1.06±0.02, 1.03±0.01,0.95±0.01 seconds main effect of group ($F(1,70) = 5.36$, $p = 0.02$).

Our hypothesis was that DBS induced modulation of GPI would enhance exploratory choices. We defined the probability of making an exploratory choice, 'P(Explore)', as the proportion of choices made in a session where the patient chose the low-value option as in an e-greedy or softmax choice rule [11]. This categorisation relies on estimates of the expected values of the available two options for low-value, exploratory choices to be defined. To identify these, we fitted the patients' choices and DT's using the RLDDM [44]. This model represents a fusion of two traditionally separate models which combine the process of iterative updating the value of expectations (Q-value Figure 2A) using the delta learning rule [46] with a choice rule (substituted for the traditional softmax) based on the sequential sampling mechanism of the Drift Diffusion Model [47].

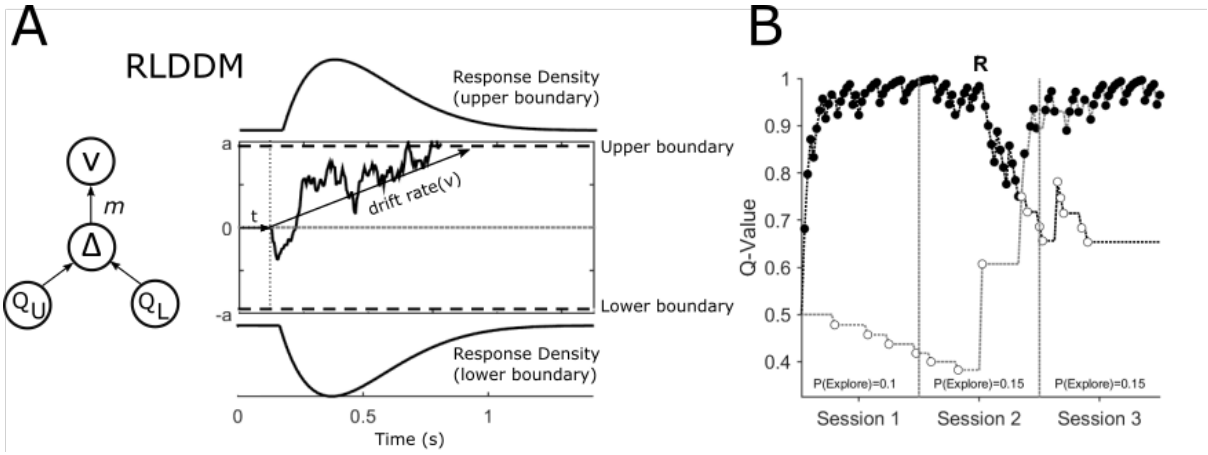

Figure 2: (A) Main features of the Reinforcement Learning Drift Diffusion Model (RLDDM). (**A**) The accumulation of evidence begins at a starting point at 0. The non-decision time is represented by t. Evidence accumulation is represented by a

sample path with added Gaussian noise and is gathered until a decision boundary (*a* and *-a*) is reached and a response is initiated. The drift rate *v* determines the rate at which this evidence is accumulated. The extent to which the difference in the expected value of the options of the upper (Q_U) and lower response boundaries (Q_L) modifies the drift rate is determined by the drift rate scaling parameter, *m*. (B) An example of an individual patient's choices (OFF-DBS, Task Block 1) across the three sessions of the task with the expected value of choices represented by the upper Q_U and lower Q_L decision boundaries in black and grey respectively. Closed circles represent "exploitative" choices where the choice with the highest expected value was chosen. Open circles are "exploratory", representing the choice of the lower value of the two options. The change in value of the two choices halfway through session 2 of the task reflects the reversal, 'R', in outcome probabilities. The probability of exploring P(Explore) is the total number of choices made for the option with the lower expected value divided by the number of choices made in the session.

The RLDDM model (Figure 2A) includes four main parameters of interest (free parameters). First, the drift rate, *v,* reflects the average speed with which the decision process approaches the response boundaries *a* (which take on positive and negative values to represent the two choices in the task). Because the *drift rate* on any one trial is proportionate to the difference in the expected value of the two choices (Figure 2A Q_U-Q_L) the scaling parameter, m, determines the extent to which the diffusion process is weighed by the difference in values of the two options. It is analogous to, and closely related to the beta parameter (inverse temperature function) which governs the explore-exploit trade-off in a softmax choice rule [44]. The expected values for the choices are in turn derived from the delta learning rule α *(R-Q) with either a single learning rate, α , or a model variant with separate learning rates for positive, $\alpha_+$and negative α_-prediction errors. The nondecision time parameter, *t*, captures the time taken by stimulus encoding and motor processes. The model fit, as measured by the Deviance Information Criterion (DIC) [48] indicated that the RLDDM with separate learning rates provided a better fit to the data compared to the model with a single learning rate (Patients: Dual learning rate DIC = 4344.98; single learning rate DIC = 4436.27; HC: Dual learning rate DIC = 3053.82; single learning rate DIC = 3056.00). Accordingly, all subsequent results reported are obtained using the dual learning rate model.

With the RLDDM model fitted to the choices separately for the ON and OFF-DBS states, we estimated the P(Explore) values for each session in the task (Figure 2D). A fixed effects analysis of P(Explore), was performed using a two-way repeated measures ANOVA with stimulation state (ON-DBS versus OFF-DBS) and task session as fixed factors. This demonstrated main effects of session ($F(2,34) = 3.14$, $p = 0.04$) and DBS state ($F(1,34) = 4.64$, $p = 0.03$) but no interaction between DBS state and the session of the task on the degree of exploration $F(2,34) = 0.77$, $p = 0.46$. In the OFF-DBS state the average P(Explore) across the sessions was 0.13 ±0.03 which increased to 0.2±0.04 when the patients performed the task with the DBS stimulator switched ON (Figure 3C). In Figure 3B & D the average P(Explore) values and reward performance for each session from fifty synthetic data sets is overlayed with the same values derived experimentally. These simulated choices were generated from RLDDM using the individual parameter estimates for each subject in the ON and OFF-DBS states from the tasks. This generated data was able to reproduce both the enhancing effect of DBS on P(Explore) and the preservation of reward learning performance in the ON and OFF-DBS states.

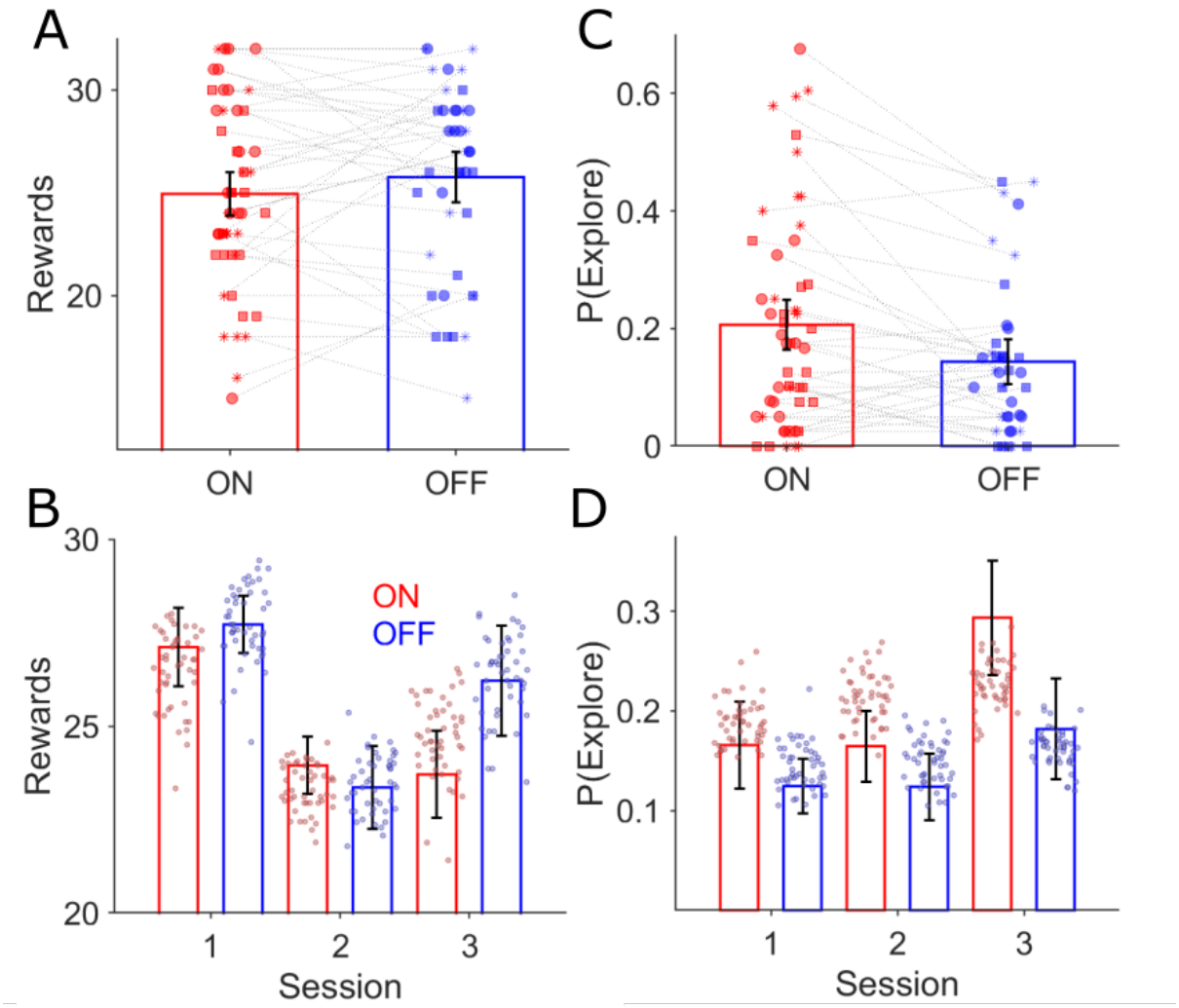

Figure 3 - Behavioural effects of GPI DBS on task performance. There was no difference in the mean rewards (P>0.05) won across the three task sessions for the ON (red bars) and OFF-DBS (blue) conditions (**A**) and (**B**). Each symbol represents a subject's number of rewards won for that session, with sessions one, two and three represented by the circle, square and asterix symbols. During ON-DBS testing the probability of exploring the lower value choice was significantly greater (p<0.05) (**C**). Behavioural performance is plotted in (**B**) (number of rewards) and (**D**) (probability of exploring the lower value choice) for both DBS conditions with the superimposed scatter plots of the 50 simulated experiments generated using the RLDDM fitted to the ON and OFF experimental choices.

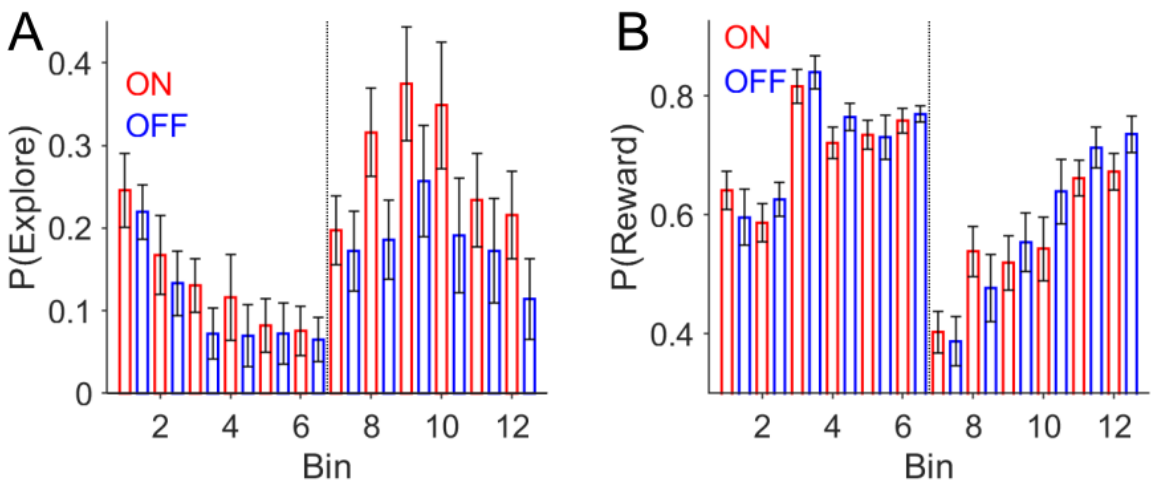

Figure 3 - Figure Supplement 1: Behavioural effects of GPI DBS on task performance peri- contingency reversal. (**A**) Probability of making an exploratory choice P(Explore), plotted as a function of time across twelve 10-trial bins. P(Explore) values are expressed as mean±s.e.m. with choices ON-DBS in red and OFF-DBS in blue. The vertical dotted line represents the point of contingency reversal in the task (midway through the second experimental session). (**B**) The probability of a receiving a reward 'P(reward)' which is the sum of the number of rewards in bin divided by the number of choices.

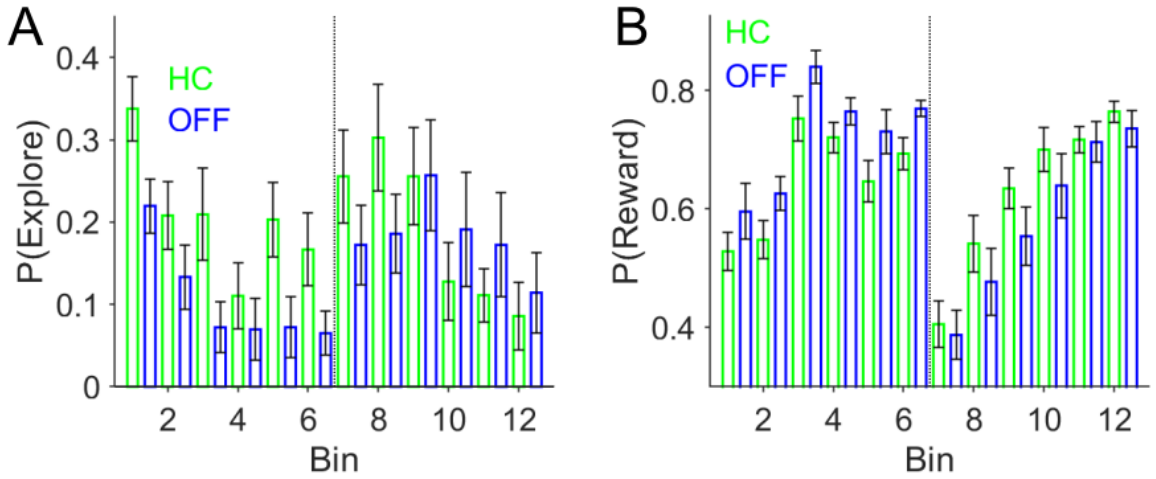

Figure 3 - Figure Supplement 2: Behavioural performance in Healthy Controls (HC) compared to patients in the DBS-OFF condition. (**A**) Probability of making an exploratory choice P(Explore), plotted as a function of time across twelve 10-trial bins. P(Explore) values are expressed as mean±s.e.m. with HC choices in green and OFF-DBS in blue. The vertical dotted line represents the point of contingency reversal in the task (midway through the second experimental session). (**B**) The probability of a receiving a reward 'P(reward)' which is the sum of the number of rewards in bin divided by the number of choices.

This analysis across the three experimental sessions in each task block may have averaged out more nuanced effects of DBS on exploration related to the reversal switch in contingencies, midway through session 2. To address this, we re-analysed the P(Explore) across twelve 10-trial bins (Figure 3 – Figure Supplement 1A). A fixed effects analysis of P(Explore) was performed using a two-way repeated measures ANOVA with stimulation state (ON-DBS versus OFF-DBS) and task bin as fixed factors. This demonstrated main effect of bin ($F(11, 187) = 5.26$, $p < 0.001$), consistent with increased P(Explore) as an effect of contingency reversal and enhanced P(Explore) values in the ON-DBS state ($F(1,187) = 9.84$, $p = 0.001$). Despite this, there was no interaction between DBS state and a specific bin of the task on the degree of exploration ($F(11,187) = 0.51$, $p = 0.89$). In view of the marked increase in P(Explore) in the post-reversal bins 7-12 (Figure 3 - Figure Supplement 1A) we also analysed the binned P(Explore) values with a fixed effect of reversal, allowing comparison of the P(Explore) values within pre-reversal bins (1-6) with those post-reversal (7-12). This demonstrated a main effect of reversal on the P(Explore) values ($F(1,187) = 27.4$, $p < 0.001$), a main effect of DBS state ($F(1,187) = 9.35$, $p = 0.002$), but no interaction between DBS state and reversal ($F(1,187) = 2.54$, $p = 0.11$). The mean P(Explore) value in each 10 trial bin in the ON-DBS state was 0.2±0.04 and 0.14±0.05 in the OFF-DBS condition.

We defined the probability of a reward, P(Reward), as the number of rewards obtained in a 10-trial bin divided by the number of responses, to assess the performance in the task on a finer time scale. This was used rather than the absolute number of rewards (as used in the analysis of session performance above) as it was a more accurate measure of performance in a small trial bin as missed trials were

accounted for. Using P(Reward) for each subject at each bin, we found no main effect of DBS ($F(1,187) = 1.88$, $p = 0.17$), or interaction between the DBS and the task bin ($F(11,187) = 0.74$, $p = 0.7$). Again, this analysis confirmed that enhanced exploratory choices in the ON-DBS state were not associated with degradation in the ability of the patients to aquire rewards during the task (mean P(Reward) values ON-DBS were 0.63±0.04 and 0.65±0.04 OFF-DBS; Figure 3 – Figure Supplement 1B).

To place the effects of DBS in our patients into the context of known abnormal reinforcement learning behaviour in isolated dystonia [49, 50], we compared the binned P(Explore) values of the HC group, to those of the patients in the OFF-DBS condition (Figure 3 – Figure Supplement 2). The probability of the HC group making an exploratory choice was greater than the patients in the OFF-DBS condition (mean HC P(Explore) = 0.19 ±0.05), confirmed by a main effect of group ($F(1,341) = 7.34$, $p = 0.005$). By contrast, there was no difference in the HC P(Explore) values when these where compared to the ON-DBS ($F(1,385) = 1.1$, $p = 0.3$) and no effect of dystonia on the overall performance of the task, as P(reward) values in the HC group were the similar (HC; mean P(Reward) = 0.63±0.03) to those of the patients in the OFF-DBS condition ($F(1,341) = 2.48$, $= 0.2$). Overall, this analysis supported that GPI-DBS enhances exploratory choices in dystonia patients from a disease related baseline level of exploitation that is higher than age matched controls.

Finally, to ensure that the order of the two reversal tasks performed OFF and ON-DBS did not contribute in any way to this result, we ran a fixed effects analysis of P(Explore) across the three sessions with stimulation state (ON-DBS versus OFF-DBS) and task order as fixed factors. No main effect of task order on P(Explore) was evident from this analysis ($F(1,34) = 0.02$, $p=0.87$, therefore, the enhanced exploration ON-DBS was unrelated to whether the first of the two reversal tasks were performed with DBS stimulator switched ON or OFF.

Cortical functional connectivity correlates of the effect of DBS on exploration.

When the DBS enhancing effects on exploratory choices on an individual subject level was investigated, this varied from a maximum within session increase of P(Explore) of 0.34, equating to 14 additional exploratory choices out of 40 in one subject, to no influence of DBS on exploration in any session in another. We leveraged the known influence of DBS on brain networks [51] to test the hypothesis that this variance could be explained by individual differences in connectivity of the DBS electrode with cortical regions which correlated strongly with exploratory decisions.

The connectivity analysis was performed based on established methodology previously introduced to assess brain networks affected by distributed lesions or stimulation effects in neurological and psychiatric conditions [52-54]. Using the stimulation volumes of the DBS electrodes as seed regions to a group connectome derived from resting state functional connectivity of healthy volunteers [55], we derived whole-brain functional connectivity maps for each patient. Then, the correlation between functional connectivity and DBS-induced change in exploratory behaviour was calculated voxel-wise. The R-Map in Figure 4B depicts thresholded correlations between functional connectivity from DBS stimulation volumes to all brain voxels and DBS-induced exploration.

We used the difference in P(Explore) values ON minus OFF-DBS as the behavioural parameter for this analysis to normalise for between subject variation in the baseline level of exploration and to identify the super-added effect of DBS on this behaviour. For each subject, we chose the experimental session with the largest increase in DBS induced exploration. This aimed to minimise the effect of averaging this behaviour across the task as a whole, whilst allowing for individual differences in the effect of DBS across different sessions of the task not evident at a group level from the analysis of P(Explore). Spatial correlation of individual patients' DBS connectivity profiles with this R-Map significantly explained variance in DBS-induced exploration (Fig. 4A) n = 14 R^2 = 0.28 p=0.04, Permutation test: R = 0.53, p = 0.009). Supplementary Table 2 includes the details of the peak R-map values and voxel co-ordinates of the cortical regions whose connectivity with the DBS electrode predicted increasing exploration. To ensure that choosing the maximum within session difference in P(Explore) for this R-map did not arbitrarily identify network effects of pallidal neuromodulation, we performed an additional R-map analysis, averaging across the whole task and subtracting the difference in P(Explore) values in the ON minus OFF- DBS conditions (Figure 4 – Figure Supplement 1). This revealed cortical connectivity patterns with similar topography. This R-map had a poorer linear predictive value of P(Explore), R^2 = 0.17, p = 0.12) but met the permutation threshold (Permutation test: R = 0.41, p = 0.004) suggesting this was unlikely to occur by chance.

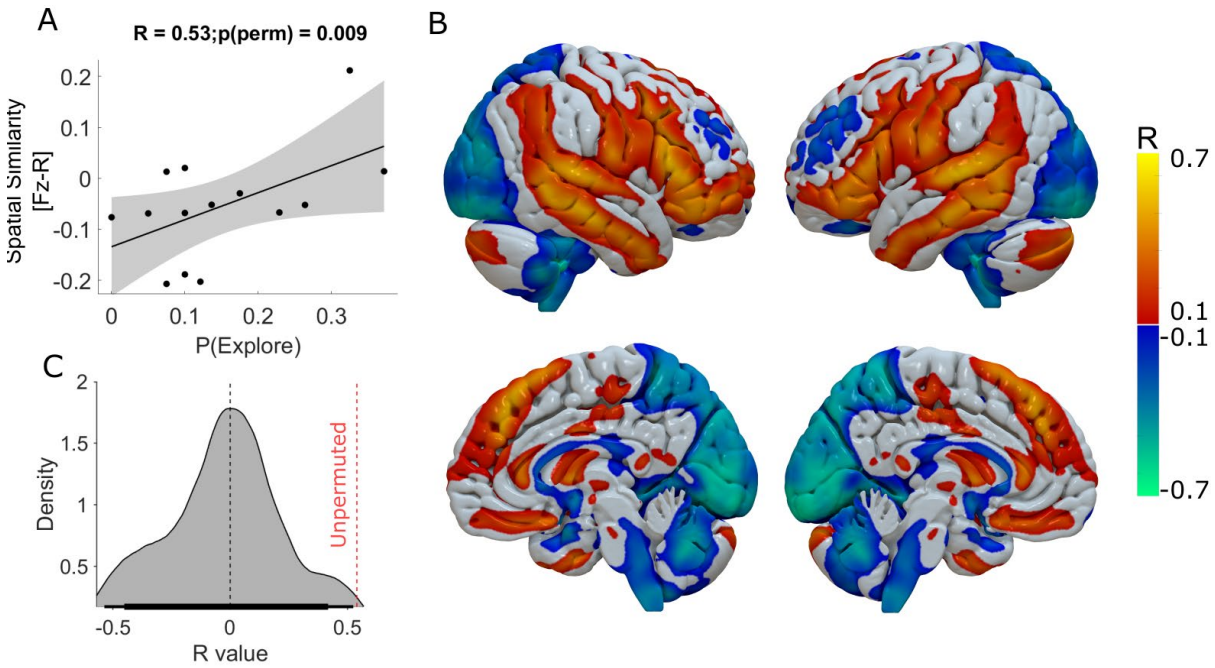

Figure 4 - Functional connectivity of DBS-Induced exploration. The whole-brain voxel-wise R-Map demonstrates the optimal functional connectivity profile for DBS-induced enhancement of exploration (B). In this analysis the maximal increase in DBS induced exploration (ON-OFF DBS) in one of the three experimental sessions was used as the regessor. Warm colours show voxels where functional connectivity to the DBS stimulation volumes was associated with greater exploration. Cool colours indicate voxels where functional connectivity to the DBS stimulation volumes was associated with lesser exploration. The more the individual functional connectivity profile matched the 'optimal' R-Map, the greater was the DBS-induced exploration (**A**) (R^2 = 0.28, P = 0.04). In (**C**) we plot the R value distribution derived from 1000 repermuted correlations between the enhancing effect DBS and the R-map. The probablity of seeing the same correlation by chance was p < 0.01 (represented by the red dashed vertical line)

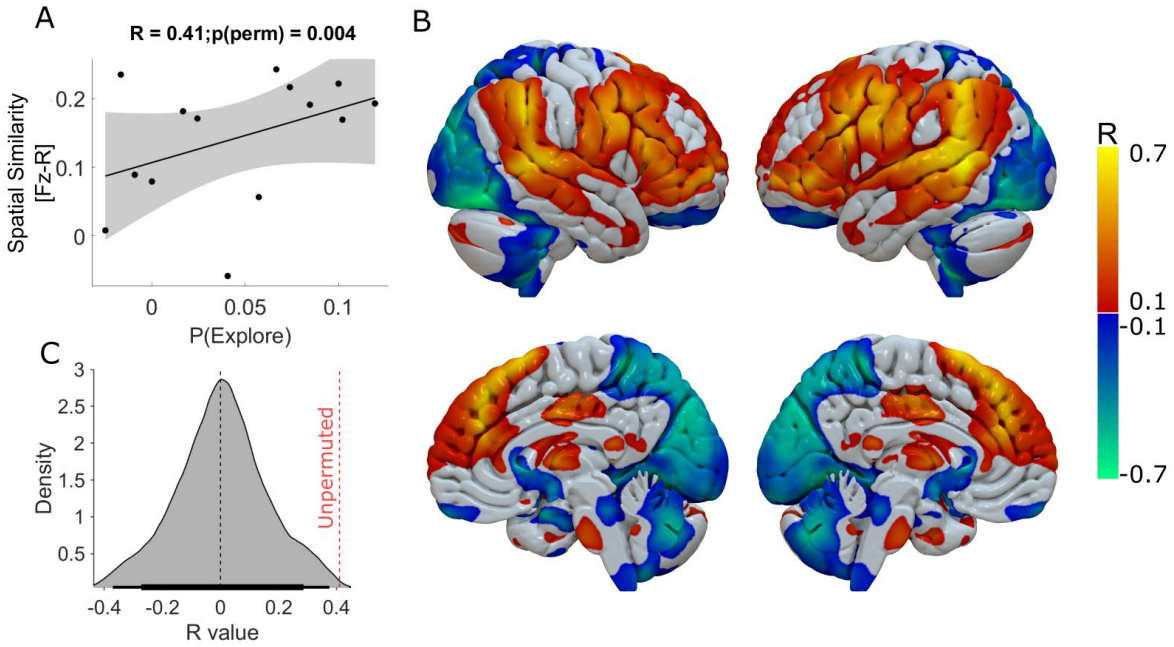

Figure 4 – Figure Supplement 1: Functional connectivity of DBS-induced exploration. The whole-brain voxel-wise R-Map demonstrates the 'optimal' functional connectivity profile for DBS-induced enhancement of exploration. **(A)** In this analysis we used the average difference in P(Explore) across the three experimental sessions in the task (ON-DBS minus OFF-DBS). (**B)** shows the 'optimal' R-Map that was associated with the highest DBS-induced exploration when assessed across the whole task. In **(C)** we plot the R value distribution derived from 1000 repermuted correlations between the enhancing DBS effect and the R-map. The probability of seeing the same correlation by chance was $p < 0.01$.

Effects of DBS on decision choices revealed by influence on RLDDM parameters: model validation and parameter recovery

We performed posterior predictive checks to test whether the dual learning rate RLDDM was a good model of the experimental data. First, we compared the observed data, in our case the performance in the reversal task and the decision time distributions, to simulated data generated by the model. The learning curves (which illustrate the probability of choosing the highest value choice) are illustrated in Figure 5A. These represent the model's generative choices for fifty simulated "experiments" with the RLDDM parameters estimated for each individual subject in both the ON and OFF-DBS conditions. The good model fit is indicated by the overlap between the generated data performance (shaded) with the observed choices (solid line) for both conditions. Furthermore, because the RLDDM includes a prediction of the decision time (DT) at which the choice was made, we also overlaid the observed DT histograms with the corresponding density for 50 generated simulations (Figure 5C). The close overlap between the observed and generated data supports that the RLDDM captured the salient features of the patients' choices and decision times whilst observing the heightened exploratory choice tendency in the ON-DBS state, as the same generative data was used for the overlaid P(Explore) simulated data in Figure 3D. As an additional check, we performed parameter recovery by re-fitting the first 5 generated data sets, plotting the model parameter estimates for the observed experimental data against the estimates derived from generated choices (Figure 6D). For all parameters, we found significant

correlations between the parameters used to generate the behaviour and the recovered parameters obtained from re-fitting the model to the generated behaviour.

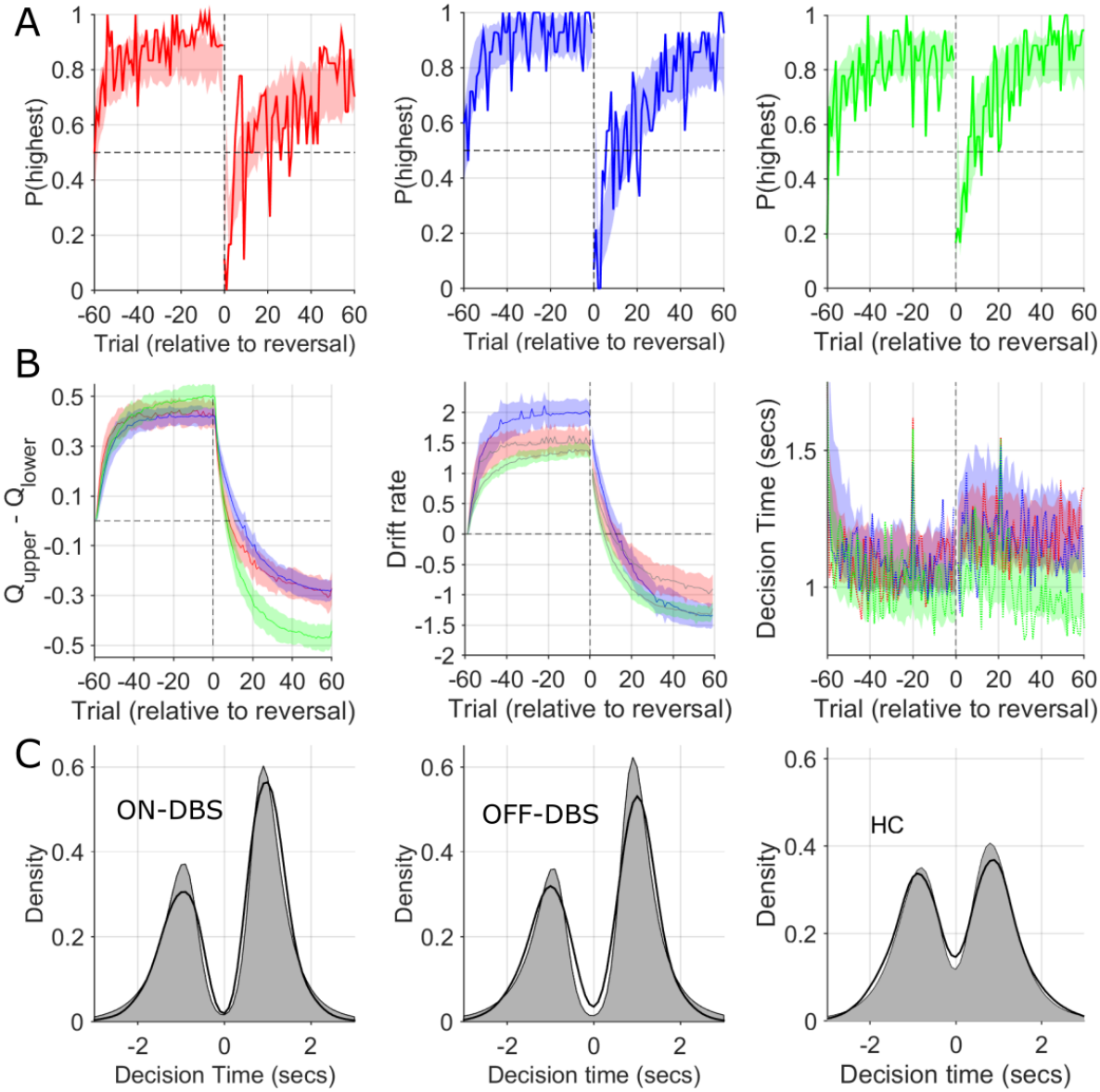

Figure 5: Observed and simulated RLDDM data. (A) The probability of choosing the highest value stimulus in the ON-DBS state is represented by the solid red line. Overlaid shaded area represents 95% confidence limits of simulated choices of the RLDDM (averaged across 50 simulated experiments). The same analysis for the OFF-DBS state is plotted in blue in the middle panel and for HC in green. This supported the interpretation of the model being a good fit to the experimental data due to the strong overlap between the synthetic data choices and those observed for both DBS conditions and in both HC and patient groups. (**B**) The difference in the expected value (Q) of the two choices in the RLDDM (mean±95% confidence limits) are represented by the solid blue (OFF-DBS), red (ON-DBS) and green (HC) lines with the shaded overlay showing the confidence limits estimated from the synthetic (simulated) data. The middle panel in **(B)** illustrates the trial-to-trial variation in the mean drift rate, *v,* across the simulations in all conditions and groups. This demonstrates a reduction in the ON-DBS drift rate consistent with the group level effect of DBS on reducing the drift rate scaling parameter, *m, (*see Figure 6 and Table 2) and the lower drift rate in the HC group. The mean experimental trial-to-trial variation in DT in each DBS condition and the HC group are plotted as a solid line with the simulated model DT's overlaid. The model captures both the within task cost of the contingency reversal on DT and the faster DT in the HC group through the task. In (**C**), thicker black lines display the observed decision time distribution across all patients and sessions in the ON-DBS (red) and OFF-DBS (blue) conditions and HC groups, with the simulated DT from

the RLDDM represented by the grey shaded regions. The decision time (DT) for choices which have an initially lower value are shown as negative. The reliability of the RLDDM in capturing the decision mechanisms in the task are supported by the overlap in observed and simulated DT distributions.

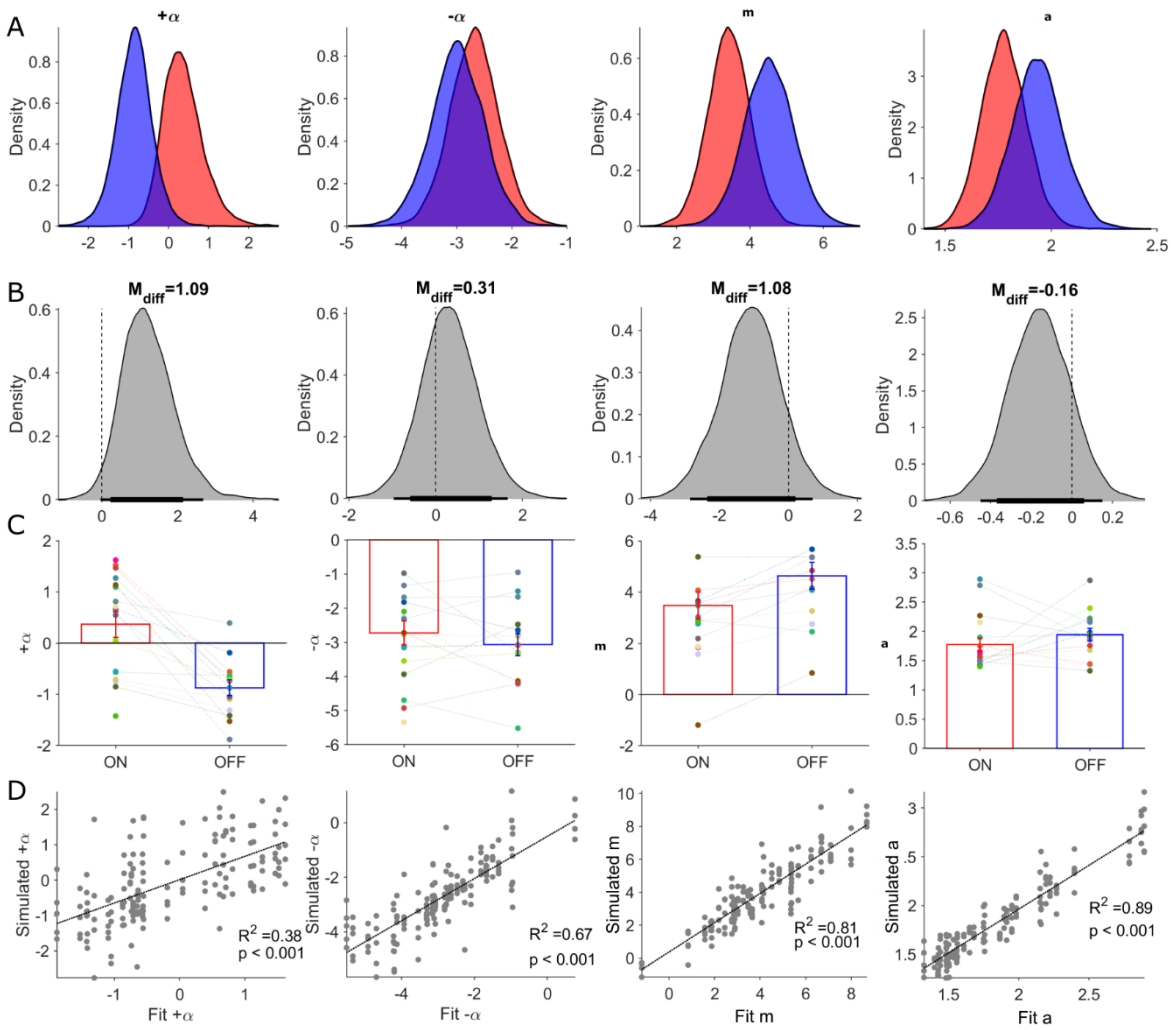

Figure 6: Posterior distributions of RLDDM parameters and parameter recovery. Columns in this figure display results for each of the estimated parameters in the RLDDM from left to right, including positive learning rate (α_+), negative learning rate (α_-), drift rate scaling parameter (m), boundary separation parameter (a). Rows, from top to bottom, correspond to (**A**) posterior distributions ON-DBS (red), OFF-DBS (blue). (**B**) Posterior distributions of differences ON-DBS versus OFF-DBS. Thick and thin horizontal bars below the distributions represent the 85% and 95% highest density intervals, respectively. **(C)** Mean ± S.E.M. parameter estimate ON-DBS (red) and OFF-DBS (blue). Each individual's parameter estimate represented by a different colour in the scatter plot for each parameter. (**D**) Results of parameter recovery analysis, plotting the estimated parameter values for the observed data against the parameter values re-estimated from simulated data. Significant (p<0.001) linear correlations were observed for all four parameters supporting successful recovery and validation of applying the RLDDM model of decision-making.

We estimated group and individual parameter values for the two learning rates (α_+ , α_-), the boundary separation parameter *(a)*, and the drift rate scaling parameter, (*m*) as dependent variables of DBS state (ON or OFF). The non-decision time, *t*,(seconds), was estimated for each subject but was assumed not to be influenced by DBS state (as DBS did not appear to affect the absolute DT in the behavioural

analysis above). Consistent with the finding of slower DTs in the patients compared to HC, the estimate for *t* in the patients was correspondingly longer (Patients = 0.57 [95% HDI 0.46, 0.69], HC = 0.42 [0.36 0.50] a difference in the posterior difference of the means which was statistically significant M_{diff} = 0.13 [0.001, 0.27].

Group level parameters of the within-subject effects of DBS were used to assess how pallidal neuromodulation influenced exploratory choices. We found no statistically significant differences between the two DBS states when the dual learning rate model was fitted to all trials across the reversal learning task (Figure 6 & Table 1). This model was therefore able to adequately capture the patient's exploratory decisions in the task and the effect of DBS (Figures 3B&D, Figure 5, Figure 6D) but provided no mechanistic insight into how pallidal neuromodulation drove an increase in these choices. Given the strong influence of the contingency reversal mid-way through session 2 of the task on exploration, (Figure 3 - Figure Supplement 1), we refitted the model separately to the pre-(1-60) and post-reversal trials (61-120). The posterior distributions of the four parameters and their posterior differences (ON minus OFF-DBS) are illustrated in Table 1- Supplement 1 and Figure 6 – Figure Supplement 1. We found a similar pattern of posterior directional effects of DBS when fitting to the whole task or separately to pre- and post-reversal. DBS led to increases in both learning rates, decreases in the boundary separation parameter *(a)* and drift rate scaling parameter (*m*). However, analysing the pre- and post-reversal trials independently, we detected the only statistically significant group level difference in the drift rate scaling parameter, *m*, in the post-reversal phase of the task; M_{diff} = - 1.71 (95% HDI [-3.5, -0.18]). This reduction in the drift rate scaling parameter *m* ON-DBS is consistent with enhanced exploration, as this parameter amplifies the influence of the difference in the expected value of the two choices (Q_u - Q_L), by driving the diffusion process to the decision boundary of the higher value option (Figure 2A).

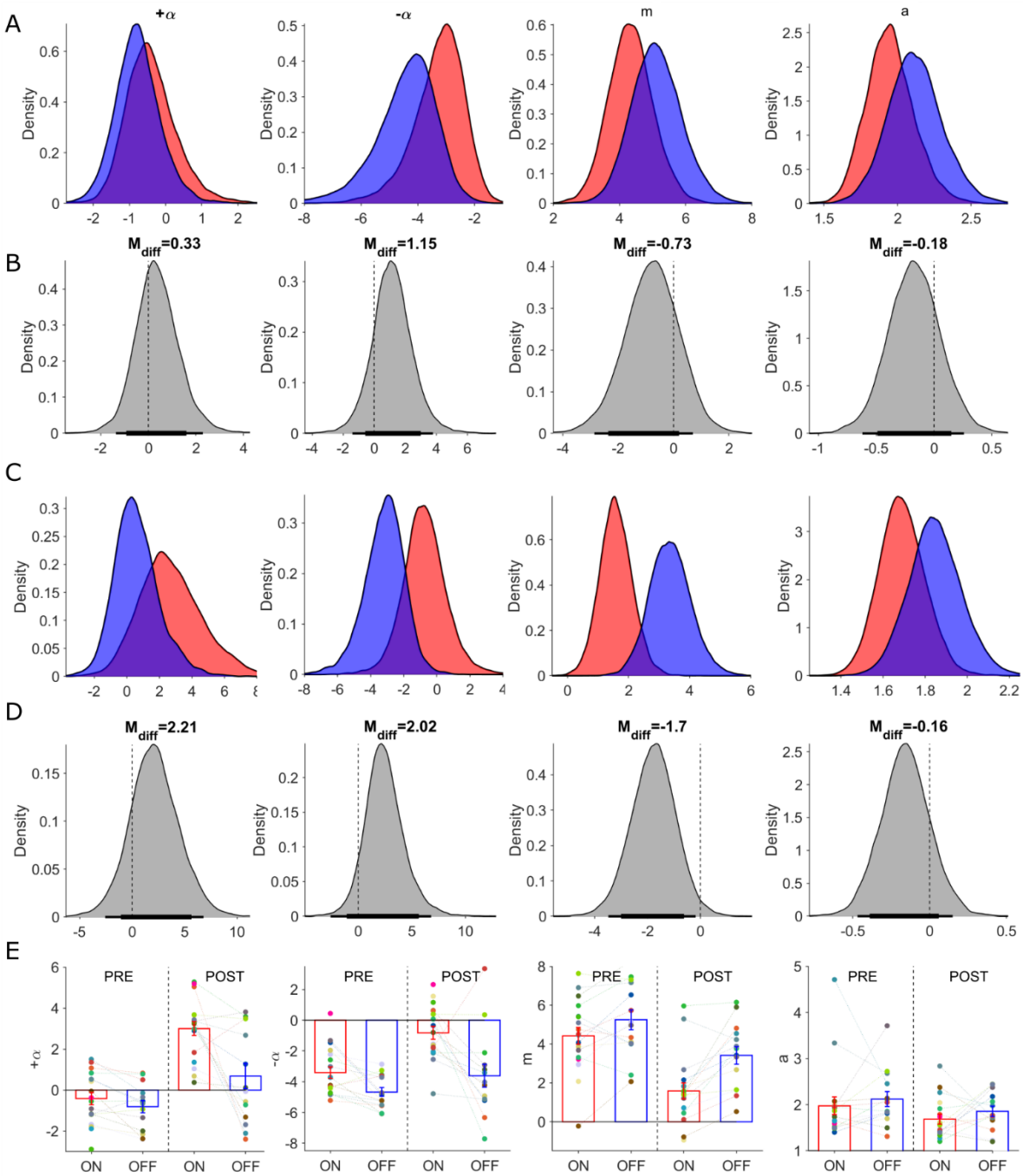

Figure 6 – Figure Supplement 1 Posterior distributions of RLDDM parameters and parameter recovery – separately fitted to pre- and post- reversal trials. Columns in this figure display results for each of the estimated parameters in the RLDDM model from left to right including, positive learning rate (α_+), negative learning rate (α_-), drift rate scaling parameter (m), boundary separation parameter (a). Parameter estimates for pre-reversal trials in rows (A) and (B) with posterior distributions ON-DBS (red), OFF-DBS (blue). (**B**) Posterior distributions of differences ON-DBS versus OFF-DBS. Thick and thin horizontal bars below the distributions represent the 85% and 95% highest density intervals, respectively. Parameter posterior distribution estimates from fitting post-reversal trials in rows (C) and posterior distributions of differences in (D) .In row (E) individual subjects parameter estimate after fitting both pre- and post reversal trials, ON (red) and OFF-DBS (blue). The group mean ± S.E.M are plotted. Each individual's parameter estimate represented by a different colour in the scatter plot for each parameter.

Comparing the RLDDM parameter estimates between the HC and patients in the OFF-DBS condition, the estimated value of, *m*, was also reduced in the HC group relative to the patients in the DBS-OFF

state (Figure 6 Figure Supplement 2, Table 1 Supplement 2) ; M_{diff} = - 1.71 (95% HDI [-3.5, -0.18]). This explains the higher proportion of exploratory choices by HC's in the task for the same reason that in the ON-DBS state exploration is heightened - this parameter proportionately scales the influence of the difference in values between the two options in on the eventual choice of decision. Patients with dystonia have previously been shown to have impaired learning from negative feedback [49, 50]. Consistent with these previous studies we also found statistically significant difference in the negative learning parameter, α_-, with a higher value HC group compared to the patients in the OFF-DBS condition M_{diff} = 1.32 (95% HDI [0.26,2.5]).

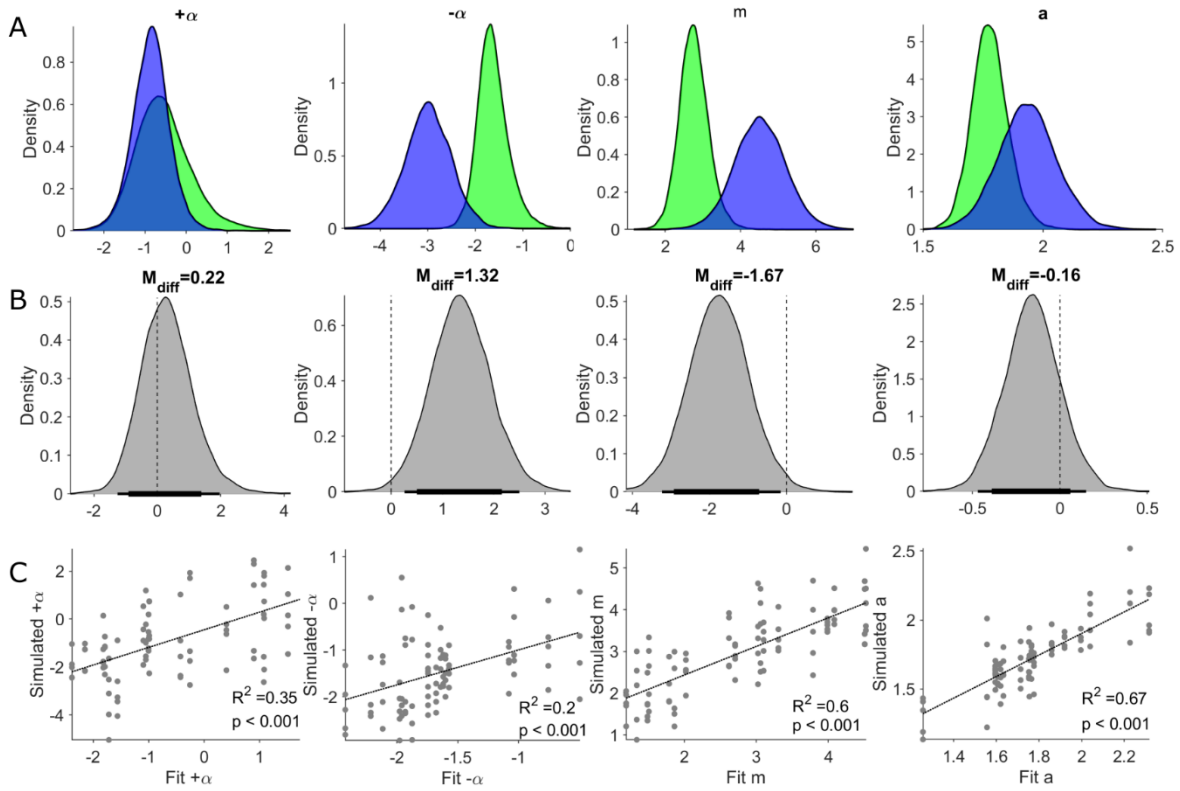

Figure 6 – Figure Supplement 2: Posterior distributions of RLDDM parameters from Healthy Controls (HC) and parameter recovery. Columns in this figure display results for each of the estimated parameters in the RLDDM model from left to right, including positive learning rate (α_+), negative learning rate (α_-), drift rate scaling parameter (m), boundary separation parameter (a). Parameter estimates in row (**A**) and with posterior distributions for HC (green) and OFF-DBS (blue). (**B**) Posterior distributions of difference between HC minus OFF-DBS. Thick and thin horizontal bars below the distributions represent the 85% and 95% highest density intervals, respectively. (**C**) Results of parameter recovery analysis, plotting the estimated parameter values for the observed data against the parameter values re-estimated from simulated data. Significant (p<0.001) linear correlations were observed for all four parameters supporting successful recovery of and validation of applying the RLDDM model of decision-making.

Table 1 - Summary of posterior distributions (RLDDM model fit to whole task)

	ON-DBS			OFF-DBS			Contrast		
Parameter	***mean***	**HDI**		***mean***	**HDI**		***mean***	**HDI**	
Boundary Separation	1.78	1.56	1.98	1.96	1.7	2.18	-0.16	-0.45	0.15
Drift rate Scaling	3.4	2.26	4.54	4.51	3.20	5.88	-1.08	-2.86	0.71
Learning rate +	0.22	-0.55	1.37	-0.85	-1.79	0.01	1.09	-0.03	2.68
Learning rate -	-2.65	-3.58	-1.8	-2.97	-3.97	-2.05	0.31	-0.97	1.66

Estimated means (m) for ON and OFF-DBS conditions as well as contrast for ON-OFF. HDI values are for 95% highest density Interval.

Table 1 - Table Supplement - 1 Summary of Posterior distributions (RLDDM model fitted separately to pre- and post-reversal trials)

	ON-DBS			OFF-DBS			Contrast		
Parameter	***mean***	**HDI**		***mean***	**HDI**		***mean***	**HDI**	
Pre-reversal									
Boundary Separation	1.94	1.63	2.27	2.07	1.76	2.50	-0.18	-0.62	0.26
Drift rate Scaling	4.28	3.07	5.72	5.00	3.70	6.66	-0.73	-2.78	1.10
Learning rate +	-0.46	-1.70	1.05	-0.82	-1.99	0.51	0.33	-1.36	2.29
Learning rate -	-2.96	-5.11	-1.67	-4.00	-6.71	-2.48	1.15	-1.40	3.79
Post-reversal									
Boundary Separation	1.66	1.48	1.90	1.83	1.60	2.10	-0.15	-0.47	0.15
Drift rate Scaling	1.5	0.55	2.63	3.4	2.03	4.69	**-1.71***	**-3.4**	**-0.18**
Learning rate +	2.16	0.80	6.78	0.34	-1.96	3.51	2.21	-2.57	6.81
Learning rate -	-0.76	-3.34	1.88	-3.09	-5.81	-0.99	2.02	-1.14	6.07

Estimated means (m) for ON and OFF-DBS conditions as well as contrast for ON-OFF. HDI values are for 95% highest density Interval. *Significant differences in the posterior distributions of the parameter estimates are highlighted in bold.

Table 1 - Table Supplement - 2 Summary of Posterior distributions (RLDDM model fitted to whole task)

	Healthy controls			Contrast (OFF-DBS)			Contrast (ON-DBS)		
Parameter	***mean***	**HDI**		***mean***	**HDI**		***mean***	**HDI**	
Boundary Separation	1.76	1.70	2.18	-0.16	-0.45	0.11	-0.03	-0.26	0.25
Drift rate Scaling	2.76	2.23	3.31	**-1.67***	**-2.93**	**-0.71**	-0.81	-2.02	0.69
Learning rate +	-0.63	-1.81	0.88	0.22	-0.91	1.39	-0.99	-2.59	0.80
Learning rate -	-1.68	-2.25	-0.95	**1.32***	**0.26**	**2.5**	0.93	-0.05	2.12

Estimated means (m) for HC as well as contrast estimates for both DBS conditions HC-OFF and HC-ON. HDI values are for 95% highest density interval. *Significant differences in the posterior distributions of the parameter estimates are highlighted in bold.

This comparison of RLDDM parameter estimates between the HC and OFF-DBS groups confirmed that greater exploratory choice tendency in HC's could be explained by a lower influence of the difference

493 in expected value of two choices. In both analyses, lower levels of exploration in HC's compared to
494 patients with dystonia and enhanced exploration by GPI neuromodulation in the patients ON-DBS, were
495 most likely explained in the RLDDM by relative reductions in the drift rate scaling parameter (*m*). In turn,
496 this meant that in both the HC group and in the patients in the ON-DBS condition, the decision to choose
497 one of the two options in each trial was much less influenced by encoding of their value. Accordingly,
498 choices were influenced proportionately more by noise intrinsic to the decision process, resulting in
499 greater proportion of random, exploratory, (and accordingly less greedy, exploitative) choices. This was
500 in turn reflected by a higher level of P(Explore) values both ON-DBS and in HC's relative to the OFF-
501 DBS condition in our behavioural analysis above.

502 **Discussion**

503 In this study we examined the role of the human pallidum (Globus Pallidus Interna) in mediating the
504 trade-off in explore-exploit decision-making. Our two predictions were confirmed: (1) Neuromodulation
505 by deep brain stimulation of the GPI increased the likelihood of patients exploring the lower value choice
506 in a 2-armed probabilistic reversal learning task; (2) DBS-induced enhanced exploration correlated with
507 the functional connectivity of the stimulation volume in the GPI to a distributed brain network including
508 frontal cortical regions identified previously in functional imaging studies of explore-exploit decision-
509 making. Furthermore, a recently proposed reinforcement learning model successfully predicted the
510 behaviour, enabling a more mechanistic interpretation of experimental results.

511 *DBS effects on GPI and their contribution to explore-exploit choice behaviour*

512 The exploration – exploitation "dilemma" is considered one of the fundamental challenges of adaptive
513 control and behaviour [1, 4]. We took advantage of the opportunity leveraged by DBS to address the
514 paucity of experimental evidence [5] relative to theoretical support for the BG role in explore-exploit
515 choice arbitration[18-20].

516 Patients and non-human primates with lesions of the GPI exhibit an inability to learn novel stimulus
517 response contingencies and random choice behaviour which prevents the acquisition of choice
518 preference or implicit learning of sequences [40, 56]. We did not observe levels of performance
519 degradation in our patients that would be comparable to complete lesioning of pallidal output.

520 Neurophysiological data support a net inhibitory effect of DBS on GPI neuron firing rates without
521 complete firing suppression at stimulation intensities comparable to those used in our patients [57]. We
522 suggest that the most likely neuromodulatory mechanism of DBS on enhanced exploration is a partial
523 inhibition of GPI firing and accordingly electrically lesioning the basal ganglia's output [58-61].

524 Previous studies have confirmed that during acute changes in GPI-DBS (comparable to the ON-OFF
525 conditions tested in our patients) a reduced inhibitory output induced by DBS in GPI leads to increased
526 excitability of the motor cortex [62, 63]. Motor cortex activity correlates negatively with decision value
527 during explore-exploit decisions [26]. This means that within a sequential sampling decision
528 mechanism, increased motor cortex activity correlates with small differences in the value of the two
529 choices and corresponding higher decision uncertainty and choice exploration. This was indeed the
530 effect predicted by the RLDDM in the ON-DBS state. A reduction in the drift-rate scaling parameter (m)

has a corresponding effect on decision value as this parameter multiplies the difference in expected value in the two choices [44]. The exploratory influence of pallidal DBS could therefore be explained by thalamocortical disinhibition, increased motor cortex excitability and its activity inversely encoding decision value.

Alternatively, rather than an influence on a downstream cortical decision mechanism, GPI-DBS may have influenced implementation of this process *within* the BG circuit. The simulations of Dunovan et. al., 2019 [64] propose that activity in the direct (striato-nigral) and indirect (striato-pallidal) BG pathways differentially encode the rate of evidence accumulation and the amount of information needed to reach decision threshold. We found that GPI-DBS reduced the drift rate scaling parameter during the post-reversal phase of the task. As this parameter determines the rate of evidence accumulation, our results would be consistent with this predicted function of the direct pathway being neuromodulated by pallidal DBS. However, because the GPI sits anatomically at the point of convergence of both direct and indirect pathways, we cannot distinguish this mechanism from a downstream influence on thalamo-cortical excitability. Future studies examining neuromodulation of a basal ganglia nucleus that lies within the indirect pathway, such as the subthalamic nucleus (STN), could be used to delineate this mechanism further, as this might be predicted to selectively influence the RLDDM boundary separation parameter which determines the decision threshold.

The basal ganglia's role in reinforcement learning (information seeking vs. uncertainty)

During the early stages of learning, firing rates of GPI neurons show transient inhibitory pauses in firing rate which encode exploratory choices. These are replaced with higher firing rates as choices become habitual and stimulus response mapping is obtained [5]. These findings are consistent with physiological changes in the direct and indirect pathways during different stages of learning. Yin et. al. [65] identified increased excitability in the striato-nigral (D1R-expressing) direct pathway early in learning with corresponding inhibition of the GPI. With subsequent consolidation of learning and habituation of the stimulus response relationship, D2R dependent increases in striato-pallidal excitability would dis-inhibit the GPI [66], increasing its excitability and promoting exploitation. Recent data from Lee and Sabatini (2021) [67] would however contradict this interpretation as they found exploration increased in optogenetically stimulated indirect pathway striato-pallidal neurons. This discrepancy might be explained by careful distinction between basal ganglia contributions to actively promoting exploratory choices during learning, versus active inhibition of choices from negative outcomes, that may be more relevant to choice extinction [68] and shifting [69].

An alternative explanation (to the idea of active exploratory information seeking) is that enhanced non-greedy choices ON-DBS represent increased random variability in the update of action values due to computational noise inherent to the learning process [8]. Examining the effects of GPI-DBS in a non-stationary 'restless' bandit task could delineate the contribution of decision noise as these models that take noise into account of the decision process have been successfully fitted to such tasks [70] .

The basal ganglia's role in reinforcement learning (learning from previous outcomes)

568 We found no effect of DBS on the learning rate to either positive or negative outcomes in the task. This
569 is surprising given evidence from Local Field Potential (LFP) recordings from GPI-DBS electrodes which
570 demonstrate that these encode choice outcome [71]. Reward Prediction Error (RPE) encoding in LFP's
571 from GPI electrodes have also been identified and are most marked during exploratory decisions and
572 absent when choices are exploitative [72]. Pallidal error-related activity has been identified from LFP
573 recordings. This proceeds the cortical error-related negativity (ERN), which is, in turn, closely related to
574 the negative prediction error signal [73, 74]. A plausible *a priori* mechanism for increased exploration
575 would be DBS induced blunting of these learning signals. The absence of any effect of DBS on the
576 RLDDM learning rate parameters results would suggest that outcome signalling including RPE-like and
577 ERN signals in the GPI are relatively unaffected by DBS. Therefore, rather than modify intrinsic
578 encoding of RPE-like signals or through distal connections with other brainstem structures, [75] DBS
579 enhanced exploration appears to be mediated by a neuromodulation of the decision process, consistent
580 with theoretical accounts proposing the basal ganglia's excitability as an arbitrator of action
581 selection[18, 19, 64].An equally parsimonious explanation for the absence of any impact of DBS on the
582 learning rate, is that by targeting of the (dorsal) motor segment of GPI, there was no effect on the
583 excitability of the ventrally located GPI – lateral habenula pathway [75] which facilitates negative
584 prediction error signalling in the Ventral Tegmental Area (VTA).

585 *Integrating the results in a network perspective*

586 The postero-lateral portion of the GPI targeted in the group of patients studied here is chosen in routine
587 clinical practice because of its position within the sensorimotor loop of the basal ganglia [76] to relieve
588 motor symptoms of dystonia. It is noteworthy that despite evidence for anatomically segmented regions
589 of the GPI (based on white matter connectivity [32]) stimulating the sensorimotor pallidum in this study
590 could alter decision-making and therefore modify "cognitive" function. At first glance, this result might
591 appear incongruous with its established role in motor control. However, our functional connectivity
592 analysis supports the conclusion that sensorimotor GPI function extends beyond that traditionally
593 attributed to it on the basis of anatomical connections. The finding of a neuromodulatory influence of
594 sensorimotor pallidum on decision-making is consistent with fMRI activation of pre-and post-central
595 gyri, during exploratory choices [2] and the encoding of decision value by motor cortex [77].

596 Our connectivity analysis is noteworthy for the extent of the connectivity pattern across much of the
597 brain. The largest within session effect of DBS was seen in patients where the electrode stimulation
598 volume shared connectivity with brain regions which would be considered functionally relevant to our
599 task (e.g. pre-frontal regions). Equally, this connectivity profile included regions with little to no
600 recognised functional role in reward-reversal learning. Therefore, caution is required in interpreting the
601 anatomical detail of this network given the limitations of using imaging data with no individualised
602 anatomical specificity. On a more general level, this analysis provides support to the idea that discrete
603 functions in local brain networks (in our case the GPI), are integrated into a broader scale whole-brain
604 network. This leads to co-variation in neural activity between regions which are not conventionally
605 attributed a functional role in a specific motor or sensory act [78] and might explain how DBS exerts
606 influence on remote areas beyond the stimulation site [51]. For instance, such as modulation of

distributed networks through subthalamic DBS affects motor learning and risk-taking behavior in Parkinson's disease patients [54, 79].

Clinical relevance - apathy.

From a clinical perspective, these findings may be relevant to understanding the mechanisms which worsen apathy in Parkinson's disease (PD) following DBS targeting the basal ganglia [80]. Patients with PD-Apathy tend to preferentially choose high value outcomes over options with lower payouts[81]. In the ON-DBS state, heightened exploratory choices observed in our data would lead to poorer encoding of the difference in value of decisions and incentivise only actions which lead to large differences in expected outcomes. This assumes that the mechanism of DBS induced apathy is analogous to apathy acquired from lesions of the bilateral globus pallidus, where the basal ganglias output is nullified, and a decrease in the signal to noise ratio that encodes value to incentivise action has been proposed [82].

Study limitations

Patients with primary forms of dystonia who have not undergone DBS exhibit abnormal reinforcement learning [35, 37, 50]. Comparison of decision making between our patients and a healthy control group replicated previously described RL abnormal bias towards explotative choices in these patients [35, 37, 50]. It seems likely that the observed enhancement of exploration by pallidal neuromodulation of GPI is from a high level of exploitaiton that is intrinsic to an imbalanced basal ganglia circuit related to the disease process in dystonia patients. We therefore cannot discount the possibility that DBS enhanced exploration was a consequence of neuromodulating a pathologically biased decision making circuit. Given the previous experimental [5] and theoretical [19] support which motivated testing the hypothesis of this study, it seems likely that the same function is subserved by a physiologically intact basal ganglia circuit. Furthermore, allowing for the uniqueness of this clinical indication for DBS in this patient group, these interpretational limitations are offset by the infrequent opportunities to study the role of this brain region in human explore-exploit decision-making.

The use of normative connectomes instead of patient specific data enhance the signal to noise ratio and image quality [51] but prevents individual quantification of connectivity strength in relation to the behavioural effect of DBS. Lastly, computational models account for a heuristic explanation of complex and dynamic neural networks and serve mainly as a theoretical support for the observed experimental results. The transferability of computationally inferred mechanisms remains a limitation of this study.

Conclusions

A recent increase in interest in the explore-exploit dilemma has significantly advanced understanding of the functional neuroanatomy, with computationally demanding cortical algorithms resolving multiple-alternative decision problems [83, 84]. The experimental results presented here demonstrate that a understanding of a unified account of the brain's approach to adaptive behavioural control benefits from inclusion of subcortical circuits. Future research should aim to delineate the differential contributions of

643 subcortical and cortical circuits to explore-exploit decision-making as well as their interaction, leading
644 to insights into disordered decision-making in psychiatric and neurologic conditions [85].

645 **Materials and Methods**

646 **Participants**

647 Nineteen patients with isolated dystonia (14f, 59.79±1.93 years old; mean ± S.E.M) who had undergone
648 Deep Brain Stimulation (DBS) surgery targeting the bilateral Globus Pallidus interna (GPI) were enrolled
649 in the study. One participant was unable to complete more than ~50% of trials in the ON-DBS state and
650 was excluded from the final analysis. Three of the remaining 18 patients were able to complete reversal-
651 learning task in the ON-DBS state but were unable to tolerate testing in the OFF-DBS state. One
652 patient's OFF-DBS testing file was inadvertently overwritten and was therefore not available for
653 analysis. For more clinical details of the patients and specifications of the inserted electrodes please
654 see supplementary Table 1.

655 Eighteen healthy controls (14f, 57.31±1.13 years old) with no previous diagnosis of a neurological or
656 psychiatric disorder were enrolled.

657 In accordance with the declaration of Helsinki, participants gave written informed consent to participate
658 in the study, which was approved by the local ethics committee (Charité – Universitätsmedizin Berlin,
659 EA1/179/20).

660 **Experimental Task**

661 We used a modified version of a two-choice reward reversal-learning task based on Pessiglione et. al.,
662 (2006) [86]. The task was presented on a laptop screen using Psychtoolbox v3.0 [87] running on
663 MATLAB (R2019, The MathWorks, Natick, MA, USA). The laptop was positioned in front of the patient
664 so that responses could be made on the keyboard. Participants were instructed to try to win as many
665 "vouchers" as possible. Printed screen text was in German. At the beginning of each trial a fixation
666 cross presented at the centre of the screen indicated that a new trial had begun. A pair of fractal images
667 were presented and subjects were expected to indicate their choice by a keyboard button press. The
668 order of the fractals was randomly assigned to either the right or left of the fixation cross. In the event
669 that they did not make a response within 3 seconds the fixation cross was re-presented and a new trial
670 began. A choice was highlighted for 0.5 seconds by a red circle around the chosen fractal. The choice
671 screen was followed by feedback screen for a further 1 second. Feedback consisted of either – "you
672 win" or a neutral feedback condition with a screen with the words "nothing". Patients performed three
673 sessions of 40 trials with 3-4 minute breaks between sessions to improve task compliance. The visual
674 stimuli were associated with a fixed probability of rewarding outcome of 80:20%. Reward contingencies
675 were reversed after 60 trial presentations (i.e. the image associated with a 80% reward probability in
676 the first 60 trials was associated with 20% reward probability in the last 60 trials and vice-versa).
677 Participants were not informed about the existence of a contingency reversal. A short training session
678 of 10 trials with a novel pair of fractals was performed before formal testing began to familiarise the
679 patients with the task.

The initial DBS condition was randomized. Patients repeated the same task OFF-DBS in the opposite DBS state after a 20min break. The fractal images differed between both task versions to avoid learning effects.

Reinforcement learning Drift Diffusion model

The RLDDM model consists of a hybrid of the sequential sampling model choice rule based on the DDM with the drift rate, v modified by the difference in the expected values of the two choices. The DDM calculates the likelihood of the decision time (DT) of choice, x , on trial, t ,with the Weiner first-passage time (WFPT) distribution;

$$DT(x) \sim WFPT[a, T_{er}, z, v(t)], \qquad (1)$$

The non-decision time and bias (starting point) of the diffusion process are represented by T_{er} and z. The boundary separation parameter a defines the point at which the diffusion process reaches the decision threshold and a response is estimated. The expected values , Q_{upper} and $Q_{lower}$of the choices represented by the upper and lower response boundaries, are estimated for each choice using the temporal difference learning rule[46] ;

$$Q_t = Q_{t-1} + \alpha * (R - Q_{t-1}), \qquad (2)$$

Where α is the learning rate and R is the outcome of the choice on trial *t.* We fitted two variations of the RLDDM model, the first with a single learning rate, the second with dual learning rates, where the Q values were modified separately by a learning rate α^+ that multiplied the positive prediction errors (i.e. outcomes that were better than expected) and by α^-, which multiplied negative prediction errors (when outcomes where worse than expected). The drift rate on each trial is then influenced by the difference in the expected values in the two choices by the drift rate scaling parameter m , where,

$$v(t) = \left[Q_{upper}(t) - Q_{lower}(t)\right] * m. \qquad (3)$$

Given our hypothesis and prior knowledge that the learning rate, scaling parameter and boundary separation parameters on explore-exploit choices [44] we estimated values of T_{er} and z with the *a priori* assumption of no effect of DBS and estimated values of α^+, α^-, m and a as dependant variables for both on and OFF-DBS states.

Model parameter estimation was performed using the Hierarchical Bayesian framework implemented in Python (3.6.12) which uses the Hierarchical estimation of DDM (HDDM) module (0.8.0) https://github.com/hddm-devs/hddm [88]. All models were run with four chains with 7,000 burn-in samples and 15000 posterior samples each. Convergence between the Markov chain Monte Carlo (MCMC) chains was assessed using the Gelman Rubin statistic [89]. For all estimates including those for parameter recovery the Gelman Rubin values were all less than 1.1 indicating successful convergence.

715 Posterior distributions in the differences in the parameter estimates were estimated by subtracting the
716 posterior estimates for each parameter in the ON and OFF-DBS states. The posterior means and
717 highest density intervals (HDI) were calculated using the 'bayestestR' and 'tidybays' packages in R
718 version 4.1.1 http://www.r-project.org [90, 91].

719 Mapping of the influence of the difference parameters on the exploratory choice preference in the
720 RLDDM model was performed by simulating 20 synthetic data sets at each parameter combination and
721 estimating the P(Explore) values. Parameter space mapping was performed for the positive learning
722 rate, drift rate scaling and boundary separation parameters (α^{+}, m and a). The parameter values for
723 each simulation where centred around the OFF state group mean estimates with each combination of
724 parameters derived from a range across of values relevant to the parameter space estimated from fitting
725 the experimental data.

726 **Lead localization and connectivity analysis**

727 DBS Electrodes were localized for 14/19 of the patients whom both testing data was available ON and
728 OFF-DBS using the software "Lead-DBS" V2.5 [92] www.lead-dbs.org as described [54, 93] and
729 mapped in MNI template space. Stimulation volumes were modelled based on SimBio-FieldTrip pipeline
730 incorporated in Lead DBS [34]. Stimulation volume modelling was informed by DBS parameters used
731 to control patients' symptoms during behavioral task (Supplementary Table 1). Each patient's
732 stimulation volumes were then used as seed regions to estimate individual functional connectivity profile
733 maps using an openly available group connectome derived from resting state functional connectivity
734 MRI images of 1000 neurologically healthy volunteers [94]. Then, voxel-wise correlations of individual
735 connectivity to DBS-induced changes in the probability of exploring the lower value choice "P(Explore)"
736 were calculated and visualized as R-maps (Figure 4B) after thresholding R [51, 54, 93]. The DBS
737 induced change in P(Explore) was defined as the maximum increase in P(Explore) ON-DBS minus
738 P(Explore) OFF-DBS in one of the three sessions. The similarity between the "ideal" connectivity profile
739 of this R-map (Figure 4B) and the connectivity profile of each individual patient was mathematically
740 assessed by calculating spatial correlation coefficients between the R-map and the individual non-
741 behavioural connectivity maps [51, 95]. Finally, the predictive potential of the DBS connectivity maps
742 for the increase in exploration was estimated by correlating the individuals increase in P(Explore) with
743 DBS to the spatial correlation coefficient of each patient (Figure 4A). A permutation distribution was
744 constructed from the same analysis performed 1000 times and the correlation coefficient re-estimated
745 with the P(Explore) values for each subject randomly re-ordered.

746 **Generating a heat-map of stimulation volumes**

747 This section describes a method to highlight the subcortical cluster stimulated by most of stimulation
748 volumes in our cohort and its relation to GPi/GPe nuclei (extracted from the DISTAL atlas[96]). This is
749 important to explore the extent of stimulation and possible differential modulation of the basal ganglia
750 nuclei in the vicinity of the stimulated area. The DISTAL atlas is a probabilistic atlas (which means it
751 offers voxel-wise probability that a voxel is belonging to a specific structure). Therefore, we first
752 thresholded bilateral GPi and GPe images to include only voxels with 50% probability of belonging to

the GPi or the GPe and extracted binary masks of them. Later, stimulation volumes were extracted and saved as binary Nifty images as has been described in the main manuscript method section. We then overlapped all binary stimulation volumes (n = 14) to extract a total stimulation volume representative of the full cohort. The latter has been explored and depicted as a 3D volume in relation to the GPi/GPe (Supplementary Fig. 1A) to visualize the extent of overlap with each of these structures. Next, we summed up the number of stimulation volumes contributing to each voxel in the total stimulation volume. This helped extracting an n-map (or heat-map) which can illustrate the frequency with which each voxel is being stimulated by stimulation volumes in the cohort. As a final step, we have calculated the extent of overlap between the heat-map and the GPi or GPe nuclei as the weighted sum of overlapping voxels. This means that we summed the values of the heat-map intersecting voxels with the binary mask of each of GPi or GPe nucleus. The latter method ensures that the voxel-wise frequency information would not be lost compared to simple binary sum of the overlap of total stimulation volume with GPi/GPe binary masks. All analyses mentioned in this section were performed in a grid of $0.5 \times 0.5 \times 0.5$ mm resolution using bihemispheric information in respect of GPi/GPe or stimulation volumes (since all patients have been bilaterally stimulated).

Behavioural data analysis

Prior to statistical analyses, trials without responses (errors of omission) were excluded (ON-DBS mean 1.06±2.26 (range 0-8) trials, OFF-DBS 1.78±2.5 (range 0 - 7 trials). Accordingly, DT and task performance (number of rewards won in a session) were analyzed using ANOVA with the MATLAB function *anovan*. In case of significant interactions, post-hoc tests were conducted using paired samples t-tests. Normality assumptions were tested using Kolmogorov Smirnov tests (all $P > 0.05$). All results are reported as mean values ±S.E.M.

Acknowledgements

TPG was funded by a NRS Fellowship from the Chief Scientist Office (Scotland). Additional funding for this research was provided by the University of Dundee/NHS Tayside Movement disorders research Endowment Fund.A.L. d e A. Marcelino is a fellow in the BIH Charité Junior Clinician Scientist Program funded by Stiftung Charité.

Competing interests

TPG has no competing interests.
ALAM has no competing interests.
JDS has no competing interests.
OG has no competing interests.
BA has no competing interests.
AAK has received from honoraria from Boston Scientific, Medtronic and Teva.

Data availability

Raw choice and reaction time data, computational model parameter estimates, simulated data and r-maps from connectivity analysis are available via the Open Science Framework https://osf.io/fs36g/

References

1. Mehlhorn, K., et al., *Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures.* Decision, 2015. **2**(3): p. 191.
2. Chakroun, K., et al., *Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making.* Elife, 2020. **9**.
3. Wilson, R.C., et al., *Humans use directed and random exploration to solve the explore–exploit dilemma.* Journal of Experimental Psychology: General, 2014. **143**(6): p. 2074.
4. Cohen, J.D., S.M. McClure, and A.J. Yu, *Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2007. **362**(1481): p. 933-942.
5. Sheth, S.A., et al., *Basal ganglia neurons dynamically facilitate exploration during associative learning.* Journal of Neuroscience, 2011. **31**(13): p. 4878-4885.
6. Herrnstein, R.J., *On the law of effect 1.* Journal of the experimental analysis of behavior, 1970. **13**(2): p. 243-266.
7. Daw, N.D., et al., *Cortical substrates for exploratory decisions in humans.* Nature, 2006. **441**(7095): p. 876-9.
8. Findling, C. and V. Wyart, *Computation noise in human learning and decision-making: origin, impact, function.* Current Opinion in Behavioral Sciences, 2021. **38**: p. 124-132.
9. Costa, V.D., et al., *Reversal learning and dopamine: a bayesian perspective.* Journal of Neuroscience, 2015. **35**(6): p. 2407-2416.
10. Izquierdo, A., et al., *The neural basis of reversal learning: an updated perspective.* Neuroscience, 2017. **345**: p. 12-26.
11. Sutton, R.S. and A.G. Barto, *Reinforcement learning: An introduction*. 2018: MIT press.
12. Findling, C., et al., *Computational noise in reward-guided learning drives behavioral variability in volatile environments.* Nat Neurosci, 2019. **22**(12): p. 2066-2077.
13. Costa, V.D., A.R. Mitz, and B.B. Averbeck, *Subcortical substrates of explore-exploit decisions in primates.* Neuron, 2019. **103**(3): p. 533-545. e5.
14. Costa, V.D., et al., *Dopamine modulates novelty seeking behavior during decision making.* Behavioral neuroscience, 2014. **128**(5): p. 556.
15. Gurney, K., T.J. Prescott, and P. Redgrave, *A computational model of action selection in the basal ganglia. I. A new functional anatomy.* Biol Cybern, 2001. **84**(6): p. 401-10.
16. Bogacz, R. and K. Gurney, *The basal ganglia and cortex implement optimal decision making between alternative actions.* Neural Comput, 2007. **19**(2): p. 442-77.
17. Suryanarayana, S.M., et al., *Roles for globus pallidus externa revealed in a computational model of action selection in the basal ganglia.* Neural Netw, 2019. **109**: p. 113-136.
18. Gilbertson, T. and D. Steele, *Tonic dopamine, uncertainty and basal ganglia action selection.* Neuroscience, 2021. **466**: p. 109-124.
19. Humphries, M.D., M. Khamassi, and K. Gurney, *Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia.* Front Neurosci, 2012. **6**: p. 9.
20. Chakravarthy, V.S., D. Joseph, and R.S. Bapi, *What do the basal ganglia do? A modeling perspective.* Biol Cybern, 2010. **103**(3): p. 237-53.
21. Hampshire, A., et al., *Dissociable roles for lateral orbitofrontal cortex and lateral prefrontal cortex during preference driven reversal learning.* Neuroimage, 2012. **59**(4): p. 4102-4112.
22. Remijnse, P.L., et al., *Neural correlates of a reversal learning task with an affectively neutral baseline: an event-related fMRI study.* Neuroimage, 2005. **26**(2): p. 609-618.

23. Cools, R., et al., *Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging.* Journal of Neuroscience, 2002. **22**(11): p. 4563-4567.
24. Ghahremani, D.G., et al., *Neural components underlying behavioral flexibility in human reversal learning.* Cerebral cortex, 2010. **20**(8): p. 1843-1852.
25. Badre, D., et al., *Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration.* Neuron, 2012. **73**(3): p. 595-607.
26. Tomov, M.S., et al., *Dissociable neural correlates of uncertainty underlie different exploration strategies.* Nat Commun, 2020. **11**(1): p. 2371.
27. Rushworth, M.F. and T.E. Behrens, *Choice, uncertainty and value in prefrontal and cingulate cortex.* Nat Neurosci, 2008. **11**(4): p. 389-97.
28. Boorman, E.D., et al., *How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action.* Neuron, 2009. **62**(5): p. 733-43.
29. Bartolo, R. and B.B. Averbeck, *Prefrontal cortex predicts state switches during reversal learning.* Neuron, 2020. **106**(6): p. 1044-1054. e4.
30. Hayden, B.Y., J.M. Pearson, and M.L. Platt, *Neuronal basis of sequential foraging decisions in a patchy environment.* Nature neuroscience, 2011. **14**(7): p. 933-939.
31. Shenhav, A., J.D. Cohen, and M.M. Botvinick, *Dorsal anterior cingulate cortex and the value of control.* Nature neuroscience, 2016. **19**(10): p. 1286-1291.
32. Draganski, B., et al., *Evidence for segregated and integrative connectivity patterns in the human Basal Ganglia.* J Neurosci, 2008. **28**(28): p. 7143-52.
33. Volkmann, J., et al., *Pallidal neurostimulation in patients with medication-refractory cervical dystonia: a randomised, sham-controlled trial.* The Lancet Neurology, 2014. **13**(9): p. 875-884.
34. Bhatia, K.P., et al., *Consensus Statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society.* Mov Disord, 2018. **33**(1): p. 75-87.
35. Gilbertson, T., M. Humphries, and J.D. Steele, *Maladaptive striatal plasticity and abnormal reward-learning in cervical dystonia.* European Journal of Neuroscience, 2019. **50**(7): p. 3191-3204.
36. Romano, R., et al., *Impaired cognitive functions in adult-onset primary cranial cervical dystonia.* Parkinsonism & related disorders, 2014. **20**(2): p. 162-165.
37. Arkadir, D., et al., *DYT1 dystonia increases risk taking in humans.* Elife, 2016. **5**: p. e14155.
38. Cleary, D.R., et al., *Deep brain stimulation entrains local neuronal firing in human globus pallidus internus.* Journal of neurophysiology, 2013. **109**(4): p. 978-987.
39. Lafreniere-Roula, M., et al., *High-frequency microstimulation in human globus pallidus and substantia nigra.* Experimental brain research, 2010. **205**(2): p. 251-261.
40. Piron, C., et al., *The globus pallidus pars interna in goal-oriented and routine behaviors: Resolving a long-standing paradox.* Movement disorders, 2016. **31**(8): p. 1146-1154.
41. Redgrave, P., et al., *Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease.* Nature Reviews Neuroscience, 2010. **11**(11): p. 760-772.
42. Daw, N.D., et al., *Model-based influences on humans' choices and striatal prediction errors.* Neuron, 2011. **69**(6): p. 1204-1215.
43. White, J.K., et al., *A neural network for information seeking.* Nat Commun, 2019. **10**(1): p. 5168.
44. Pedersen, M.L., M.J. Frank, and G. Biele, *The drift diffusion model as the choice rule in reinforcement learning.* Psychonomic bulletin & review, 2017. **24**(4): p. 1234-1251.
45. Edlow, B.L., et al., *7 Tesla MRI of the ex vivo human brain at 100 micron resolution.* Scientific data, 2019. **6**(1): p. 1-10.
46. Rescorla, R.A., *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.* Current research and theory, 1972: p. 64-99.

47. Ratcliff, R., *A theory of memory retrieval.* Psychological review, 1978. **85**(2): p. 59.
48. Spiegelhalter, D.J., et al., *Bayesian measures of model complexity and fit.* Journal of the royal statistical society: Series b (statistical methodology), 2002. **64**(4): p. 583-639.
49. Gilbertson, T., M. Humphries, and J.D. Steele, *Maladaptive striatal plasticity and abnormal reward-learning in cervical dystonia.* Eur J Neurosci, 2019. **50**(7): p. 3191-3204.
50. Gilbertson, T., D. Arkadir, and J.D. Steele, *Opposing patterns of abnormal D1 and D2 receptor dependent cortico-striatal plasticity explain increased risk taking in patients with DYT1 dystonia.* PloS one, 2020. **15**(5): p. e0226790.
51. Horn, A., et al., *Connectivity predicts deep brain stimulation outcome in P arkinson disease.* Annals of neurology, 2017. **82**(1): p. 67-78.
52. Boes, A.D., et al., *Network localization of neurological symptoms from focal brain lesions.* Brain, 2015. **138**(10): p. 3061-3075.
53. Corp, D.T., et al., *Network localization of cervical dystonia based on causal brain lesions.* Brain, 2019. **142**(6): p. 1660-1674.
54. de Almeida Marcelino, A.L., et al., *Subthalamic neuromodulation improves short-term motor learning in Parkinson's disease.* Brain, 2019. **142**(8): p. 2198-2206.
55. Thomas Yeo, B., et al., *The organization of the human cerebral cortex estimated by intrinsic functional connectivity.* Journal of neurophysiology, 2011. **106**(3): p. 1125-1165.
56. Obeso, J., et al., *What can man do without basal ganglia motor output? The effect of combined unilateral subthalamotomy and pallidotomy in a patient with Parkinson's disease.* Experimental neurology, 2009. **220**(2): p. 283-292.
57. Bar-Gad, I., et al., *Complex locking rather than complete cessation of neuronal activity in the globus pallidus of a 1-methyl-4-phenyl-1, 2, 3, 6-tetrahydropyridine-treated primate in response to pallidal microstimulation.* Journal of Neuroscience, 2004. **24**(33): p. 7410-7419.
58. Wu, Y., et al., *Does stimulation of the GPi control dyskinesia by activating inhibitory axons?* Movement disorders: official journal of the Movement Disorder Society, 2001. **16**(2): p. 208-216.
59. Dostrovsky, J., et al., *Microstimulation-induced inhibition of neuronal firing in human globus pallidus.* Journal of neurophysiology, 2000. **84**(1): p. 570-574.
60. McCairn, K.W. and R.S. Turner, *Deep brain stimulation of the globus pallidus internus in the parkinsonian primate: local entrainment and suppression of low-frequency oscillations.* Journal of neurophysiology, 2009. **101**(4): p. 1941-1960.
61. Boraud, T., et al., *High frequency stimulation of the internal Globus Pallidus (GPi) simultaneously improves parkinsonian symptoms and reduces the firing frequency of GPi neurons in the MPTP-treated monkey.* Neuroscience letters, 1996. **215**(1): p. 17-20.
62. Tisch, S., et al., *Pallidal stimulation modifies after-effects of paired associative stimulation on motor cortex excitability in primary generalised dystonia.* Experimental neurology, 2007. **206**(1): p. 80-85.
63. Kühn, A., et al., *Modulation of motor cortex excitability by pallidal stimulation in patients with severe dystonia.* Neurology, 2003. **60**(5): p. 768-774.
64. Dunovan, K., et al., *Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making.* PLoS computational biology, 2019. **15**(5): p. e1006998.
65. Yin, H.H., et al., *Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill.* Nature neuroscience, 2009. **12**(3): p. 333-341.
66. Kravitz, A.V., et al., *Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry.* Nature, 2010. **466**(7306): p. 622-626.
67. Lee, J. and B.L. Sabatini, *Striatal indirect pathway mediates exploration via collicular competition.* Nature, 2021. **599**(7886): p. 645-649.
68. Tecuapetla, F., et al., *Complementary contributions of striatal projection pathways to action initiation and execution.* Cell, 2016. **166**(3): p. 703-715.

69. Kravitz, A.V., L.D. Tye, and A.C. Kreitzer, *Distinct roles for direct and indirect pathway striatal neurons in reinforcement.* Nature neuroscience, 2012. **15**(6): p. 816-818.
70. Findling, C., et al., *Computational noise in reward-guided learning drives behavioral variability in volatile environments.* Nature neuroscience, 2019. **22**(12): p. 2066-2077.
71. Eisinger, R.S., et al., *Distinct Roles of the Human Subthalamic Nucleus and Dorsal Pallidum in Parkinson's Disease Impulsivity.* Biological Psychiatry, 2021.
72. Schroll, H., et al., *Differential contributions of the globus pallidus and ventral thalamus to stimulus–response learning in humans.* NeuroImage, 2015. **122**: p. 233-245.
73. Herrojo Ruiz, M., et al., *Involvement of human internal globus pallidus in the early modulation of cortical error-related activity.* Cerebral Cortex, 2014. **24**(6): p. 1502-1517.
74. Holroyd, C.B., et al., *Errors in reward prediction are reflected in the event-related brain potential.* Neuroreport, 2003. **14**(18): p. 2481-2484.
75. Hong, S. and O. Hikosaka, *The globus pallidus sends reward-related signals to the lateral habenula.* Neuron, 2008. **60**(4): p. 720-729.
76. Alexander, G.E., M.R. DeLong, and P.L. Strick, *Parallel organization of functionally segregated circuits linking basal ganglia and cortex.* Annual review of neuroscience, 1986. **9**(1): p. 357-381.
77. Hare, T.A., et al., *Transformation of stimulus value signals into motor commands during simple choice.* Proceedings of the National Academy of Sciences, 2011. **108**(44): p. 18120-18125.
78. Kauvar, I.V., et al., *Cortical Observation by Synchronous Multifocal Optical Sampling Reveals Widespread Population Encoding of Actions.* Neuron, 2020. **107**(2): p. 351-367 e19.
79. Irmen, F., et al., *Sensorimotor subthalamic stimulation restores risk-reward trade-off in Parkinson's disease.* Movement Disorders, 2019. **34**(3): p. 366-376.
80. Zoon, T.J., et al., *Apathy Induced by Subthalamic Nucleus Deep Brain Stimulation in Parkinson's Disease: A Meta-Analysis.* Movement disorders, 2021. **36**(2): p. 317-326.
81. Le Heron, C., et al., *Distinct effects of apathy and dopamine on effort-based decision-making in Parkinson's disease.* Brain, 2018. **141**(5): p. 1455-1469.
82. Levy, R. and B. Dubois, *Apathy and the functional anatomy of the prefrontal cortex–basal ganglia circuits.* Cerebral cortex, 2006. **16**(7): p. 916-928.
83. Gershman, S.J., *Deconstructing the human algorithms for exploration.* Cognition, 2018. **173**: p. 34-42.
84. Schulz, E. and S.J. Gershman, *The algorithmic architecture of exploration in the human brain.* Current opinion in neurobiology, 2019. **55**: p. 7-14.
85. Addicott, M.A., et al., *A primer on foraging and the explore/exploit trade-off for psychiatry research.* Neuropsychopharmacology, 2017. **42**(10): p. 1931-1939.
86. Pessiglione, M., et al., *Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans.* Nature, 2006. **442**(7106): p. 1042-1045.
87. Brainard, D.H. and S. Vision, *The psychophysics toolbox.* Spatial vision, 1997. **10**(4): p. 433-436.
88. Wiecki, T.V., I. Sofer, and M.J. Frank, *HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python.* Frontiers in neuroinformatics, 2013. **7**: p. 14.
89. Gelman, A., *Two simple examples for understanding posterior p-values whose distributions are far from uniform.* Electronic Journal of Statistics, 2013. **7**: p. 2595-2602.
90. Kay, M., *tidybayes: Tidy data and geoms for Bayesian models.* R package version, 2019. **1**(0).
91. Makowski, D., M.S. Ben-Shachar, and D. Lüdecke, *bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework.* Journal of Open Source Software, 2019. **4**(40): p. 1541.
92. Horn, A., et al., *Lead-DBS v2: Towards a comprehensive pipeline for deep brain stimulation imaging.* Neuroimage, 2019. **184**: p. 293-316.

93. Neumann, W.-J., et al., *Functional segregation of basal ganglia pathways in Parkinson's disease.* Brain, 2018. **141**(9): p. 2655-2669.
94. Holmes, A.J., et al., *Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures.* Sci Data, 2015. **2**: p. 150031.
95. Al-Fatly, B., et al., *Connectivity profile of thalamic deep brain stimulation to effectively treat essential tremor.* Brain, 2019. **142**(10): p. 3086-3098.
96. Ewert, S., et al., *Toward defining deep brain stimulation targets in MNI space: a subcortical atlas based on multimodal MRI, histology and structural connectivity.* Neuroimage, 2018. **170**: p. 271-282.

1001

Demographics					Stimulation details		TWSTRS-Score					Tsui-Score		BDI
ID	**Sex**	**Age**	**Disease Duration**	**DBS Duration**	**Parameters**	**Stim Start**	**Severity ON**	**Severity OFF**	**Impairment ON**	**Pain ON**	**Total ON**	**TSUI ON**	**TSUI OFF**	**BDI**
0	F	54	22	8	L: 10-, 3.1V; R: 1-, 3.1V; 130Hz, 90µs	OFF	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
1	F	56	12	8	L: 9-, 1.1V; R: 1-, 1.7V; 210Hz, 120µs	ON	9	15	2	3.5	14.5	2	8	2
2	F	45	5	2	L: 1-, 2-, 1.9V; R: 10-, 11-, 3.6V; 210Hz, 90µs	OFF	3	6	1	9.25	13.25	1	n.a.	17
3	F	55	8	2	L: 3-, 3mA; R: 11-, 3mA; 185Hz, 90µs	ON	7	23	0	1.75	8.75	2	10	6
4	M	67	32	8	L: 9-, 10-, 3.1V; R: 1-, 2-, 2.6V; 180Hz, 90µs	OFF	6	7	17	11.75	34.75	2	4	11
5	M	55	21	10	L: 9-, 2.2V; R: 1-, 4V; 60Hz, 90µs	OFF	2	4	0	2	4	2	2	4
6	F	61	19	5	L: K4- (30%), K5- (70%), 3.3mA; R: 12- (30%), K13- (70%); 231Hz, 120µs	ON	14	15	13	15	42	7	8	10
7	M	56	17	14	L: K9-, 2.1V; R: 1-, 2.1V; 130Hz, 90µs	ON*	6	n.a.	8	0.75	14.75	2	n.a.	n.a.
8	F	61	14	10	L: K9-, 10-, 1.5V; R: 1-, 2-, 1.8V; 180Hz, 90µs	OFF	3	4	3	0	3	1	2	1
9	F	73	11	0	L: 2-, 3-, 2.5mA; R: 10-, 11-, 2.5mA; 130Hz, 60µs	ON	16	19	n.a.	n.a.	n.a.	9	10	7
10	F	63	32	1	L: 2-, 3-, 2.9mA; R: 11-, 12-, 2.6mA, 130Hz, 90µs	ON*	13	11	15	1	29	8	8	21
11	F	68	15	11	L: 9-, 10-, 11- vs. 8+, 3.1V; R: 1-, 2-, 3- vs. 0+, 2.5V; 200Hz, 60µs	ON*	6	21	10	7	23	3	9	22

12	M	56	22	11	L: 10-, 11-, 2.8V; R: 2-, 3-, 2.1V; 60Hz, 120μs	OFF	11	12	10	12.5	33.5	7	7	16
13	F	70	37	11	L: 10-, 3.3V, 140Hz, 90μs; R: 1-, 2.8V, 140Hz, 120μs	ON*	23	n.a.	17	14.25	54.25	10	n.a.	10
14	M	43	19	2	L: 2-, 5.4mA; R: 2-, 3.3mA; 130Hz, 90μs	OFF	15	18	6	4.75	25.75	10	12	4
15	F	54	29	18	L: 5-, 6-, 3.0V; R: 1-, 2-, 3.7V; 180Hz, 90μs	ON*	5	n.a.	5	0	10	3	n.a.	20
16	F	71	17	4	L: 7-, 2.1V; R: 1-, 2.5 V; 65Hz, 60μs	ON	4	8	2	6	12	2	6	6
17	F	71	26	6	L: 9-, 3.1V; R: 0-, 3.2V; 140Hz, 90μs	OFF	4	10	1	8	13	2	3	9
18	F	59	31	3	L: 8-, 9-, 2.2V; R: 0-, 1-, 2V; 200Hz, 90μs	OFF	3	5	0	2.5	5.5	0	1	20
Mean		59.79	20.47	7.13			8.33	11.87	6.47	5.88	20.06	4.06	6.43	10.94
S.E.M.		1.93	2.03	1.14			1.37	1.64	1.49	1.24	3.52	0.80	0.93	1.71

Supplementary Table 1 TWSTRS: Toronto Western Spasmodic Torticollis Rating Scale; BDI: Beck's Depression Inventory; F: Female; M: Male; S.E.M: standard error of mean; L: left; R: right. *These patients only performed the task in one stimulation condition.

Positive Peak Coordinates X/Y/Z (Value)			*Negative Peak Coordinates X/Y/Z (Value)*		
VLPFC (BA 45)	RH	56/28/18 (0.58)	CBM	RH	32/-36/-24 (-0.64)
	LH	-56/22/12 (0.50)		LH	-30/-50/-20 (-0.60)
CBM	RH	38/-82/-26 (0.49)	OL (BA 18)	RH	36/-94/-12 (-0.62)
	LH	-48/-62/-38 (0.65)		LH	-4/-98/-12 (-0.61)
Uncus (BA 36)	RH	30/-4/-38 (0.48)	PostCG (BA 5)	RH	26/-42/72 (-0.45)
	LH	-22/4/-46 (0.41)		LH	-24/-36/64 (-0.37)
SFG (BA 8)	RH	8/34/52 (0.60)	Prec (BA 7)	RH	12/-46/64 (-0.60)
	LH	-1/34/54 (0.57)		LH	-4/-74/38 (-0.59)
DLPFC (BA 9)	RH	58/24/20 (0.59)	Midbrain	RH	8/-32/-14 (-0.51)
	LH	-56/22/20 (0.50)		LH	-8/-30/-12 (-0.49)
MTG (BA 21)	RH	60/-24/-18 (0.58)			
	LH	-66/-52/-2 (0.54)			
MedFG (BA 6, 9)	RH	4/42/30 (0.54)			
	LH	-2/42/34 (0.53)			
SMG (BA 40)	RH	60/-48/36 (0.59)			
	LH	-54/-52/34 (0.48)			
ITG (BA 21)	RH	60/-10/-28 (0.54)			
	LH	-64/-18/-24 (0.51)			
ACC (BA 32)	RH	12/38/12 (0.54)			
	LH	-14/40/8 (0.43)			
MCC (BA 24)	RH	4/-4/30 (0.45)			
	LH	-1/-6/30 (0.46)			
PreCG (BA 4)	RH	60/-12/40 (0.52)			

	LH	-64/-12/30 (0.55)
PostCG (BA 1)	RH	58/-14/44 (0.52)
	LH	-60/-14/38 (0.53)
SMA (BA 6)	RH	33/-10/65 (0.42)
	LH	-31/-10/65 (0.36)
Putamen	RH	32/-6/-10 (0.56)
	LH	-24/2/-10 (0.47)

Supplementary Table 2 Summary of connectivity (R-map) analysis ACC, anterior cingulate cortex; BA, Brodmann area; CBM, cerebellum; IFG, inferior frontal gyrus, ins., insula; ITG, inferior temporal gyrus; MCC, midcingulate cortex; MedFG, medial frontal gyrus; MFG, middle frontal gyrus; MTG, middle temporal gyrus; OG, orbital gyrus; OL, occipital lobule; PCC, posterior cingulate cortex; PreCG, precentral gyrus; Prec, precuneus; PostCG, postcentral gyrus; SFG, superior frontal gyrus; SMG, supramarginal gurys; SNr, substantia nigra; STG, superior temporal gyrus

1042