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], 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: 7587

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 both rate of evidence accumulation (drift rate) and by the amount of evidence needed to decide (boundary separation) 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 predicted by the degree of functional connectivity between the stimulating electrode and prefrontal and sensorimotor cortices. 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]. Despite the relevance of this 'dilemma' to understanding decision making in humans, its neural basis remains poorly understood [2].

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 outcome/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 [3], 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 [4]. Conversely, 'greedy' choices are made to the option with the highest value [5-7] and can be considered to exploit current knowledge of the values of the actions [8]. One possibility is that exploration represents a cognitive strategy for active information seeking, to reaffirm existing beliefs in the relative value of the choices [7-9]. Alternatively, it may represent random choice perturbations caused by limitations in brain's implementation of the decision process leading to decision noise [10]. Either way, the strategies adopted by human participants performing bandit tasks, including reversal-learning [6, 11, 12], rely on exploiting stability whilst at the time same time adapting to volatility by exploring alternatives courses of action.

Several observations support a role of the sub-cortical nuclei that make up the basal ganglia (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 [13, 14]. 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 [15]. In the context of explore-exploit dilemma, this means the BG may implement a neural 'decision filter,' with the bandwidth of this filter being modulated by neuromodulators, such as dopamine, which influence BG excitability [16]. 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 [17, 18]. Consistent with this prediction, the firing rate of GPI neurons in non-human primates encode the transition from exploratory to exploitative decision making [19]. 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 [20-23], ventromedial prefrontal (vmPFC) and dorsolateral prefrontal (DLPFC) 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[4, 24]. 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) [4, 10, 24-26] 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 or focal and segmental dystonia [33]. Patients with dystonia have normal levels of striatal dopamine, intact reward prediction error (RPE) signalling [34], and have only subtle cognitive deficits [35]. Dystonia patients with chronically implanted DBS-GPI electrodes therefore represent an 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.

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 [36, 37] we hypothesised that DBS mediated reduced GPI excitability should drive greater proportion of exploratory choices consistent with previous findings by Sheth et. al., 2011 [38]. 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. Previously, studies of the explore-exploit dilemma have utilised bandit tasks with high pay-out volatility and/or multiple choice options [4, 24]. Given the more recognised role of the BG in habitual learning of stimulus-response associations [39, 40], we purposely chose a low volatility probabilistic reversal-learning task to engage cortico-sub-cortical circuits in "model-free" learning [41]. 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 [42]. 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.

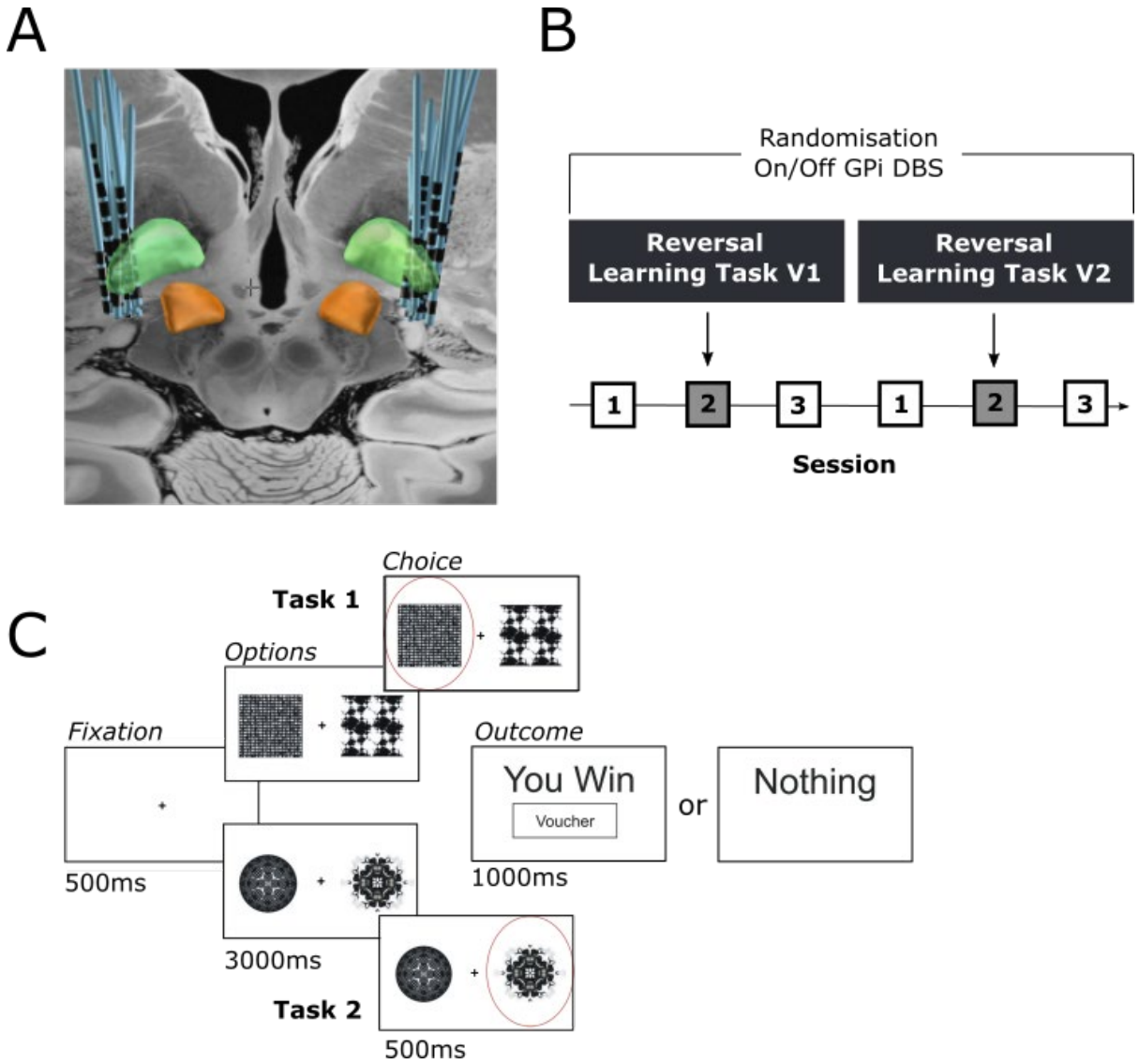

Figure 1 General overview of the study design and task. (**A**) Localization of DBS electrodes was done with Lead-DBS (www.lead-dbs.org) as previously described[43]. Electrodes are implanted within the GPI (green). The subthalamic nucleus (STN) is shown in orange. (**B**) The reversal learning task consisted 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 (V1 and V2).(C) Example of a trial in the reversal learning tasks. 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 V1 the left and cue, in Task V2 the right-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%.

Results

GPI DBS enhances exploratory choices without affecting task performance

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 electrodes turned "ON" or "OFF" in a randomised order between the two tasks. The task 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 task versions differed only by the fractals presented (Figure 1C). A fixed effects analysis of performance in the task (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 ($F(2,17) = 6.26$, $P = 0.002$, 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 ($F(1,17) = 0.9$, $p = 0.34$) or interaction between DBS and performance in a session of the task ($F(2,17) = 1.03$, $p = 0.30$) see Figure 3 A&C. 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 ($F(2,17) = 5.21$, $p = 0.005$, DTs: Sessions 1, 2 and 3 ; 1.04±0.02, 1.14±0.01,1.18±0.01 seconds) but no effect of DBS ($F(1,17) = 0.01$, $p = 0.92$ or interaction between DBS and the decision times within a session of the task ($F(2,17) = 1.27$, $p = 0.27$). This analysis confirmed that the switching their DBS stimulator to the OFF state did not affect task performance by worsening the patients symptom's.

Our hypothesis was that DBS induced modulation of GPI would enhance exploratory choices. We defined the probability of making an exploratory choice in a session, '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 [8]. 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 model [42]. This model represents a fusion of two traditionally separate models which combine the process of iterative updating the value of expectations (Q-value Figure 2) using the delta learning rule [44] with a choice rule (substituted for the traditional softmax) based on the sequential sampling mechanism of the Drift Diffusion Model [45].

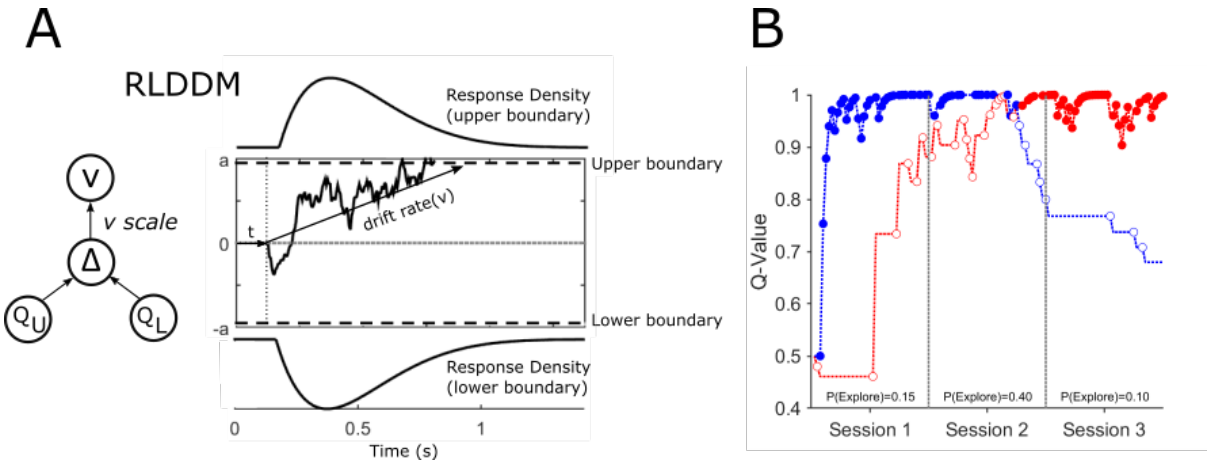

Figure 2: (A) Main features of the Reinforcement learning drift diffusion model (RLDDM). The accumulation of evidence begins at a starting point at 0. The non-decision time is represented by t. Evidence accumulation is represented by a sample path with added Gaussian noise and is gathered until a decision 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 patients choices across the three sessions of the task with the expected value of the upper Q_U and lower Q_L choices in blue and red. The 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 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 2 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. 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) [47] indicated that the RLDDM with separate learning rates provided a better fit to the data compared to the model with a single learning rate (Dual learning rate DIC = 4344.98; single learning rate DIC = 4436.27). 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 2B). 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 ($F(2,17) = 3.14$, $p = 0.04$) and DBS state ($F(1,17) = 4.64$, $p = 0.03$) but no interaction between DBS state and the session of the task on the degree of exploration $F(2,17) = 0.77$, $p = 0.46$. 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 (Figure 3A) patients performed the task with the DBS stimulator switched ON (post-hoc paired singled tailed t-test t(93) = 2.05, p = 0.02). 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.

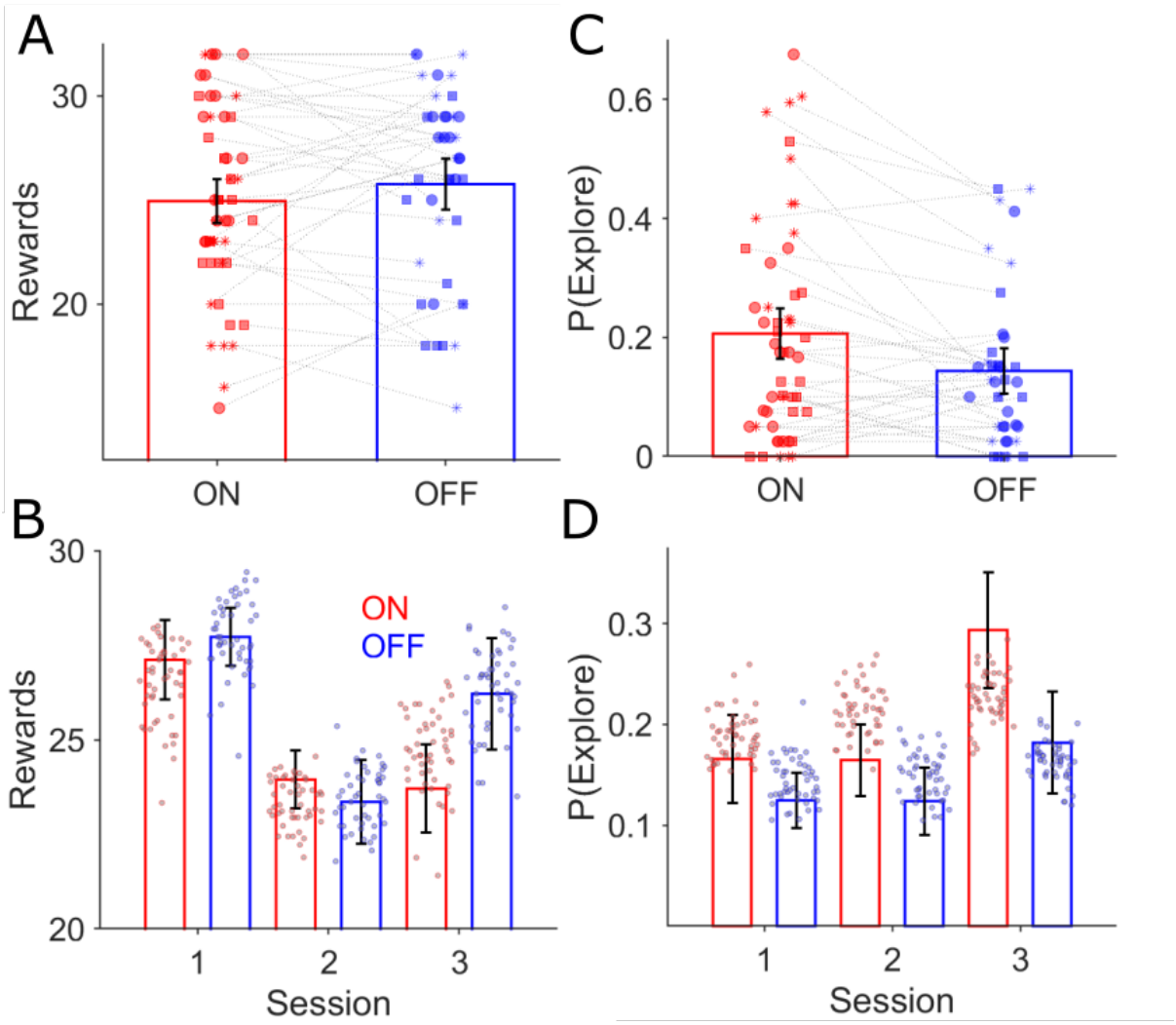

Figure 3 Behavioural effects of GPI DBS on task performance. There was no difference in the mean rewards won across the three task sessions p>0.05 for the ON (red bars) and OFF-DBS (blue) conditions (**A**) and (**B**). Each symbol represents a subject's number of rewards won for that session, with sessions 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 when ON-DBS (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

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 [48] 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. Using the stimulation volumes of the DBS electrode as seed regions to a group connectome derived from resting state functional connectivity [49], we derived whole-brain functional connectivity maps for each patient. These connectivity maps were then used to calculate a whole-brain "R-Map"(Figure 4B) that depicted correlations between individual connectivity profiles from DBS stimulation volumes and DBS-induced exploration (maximum within session increase in P(Explore) DBS-ON - P(Explore) DBS-OFF). 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.

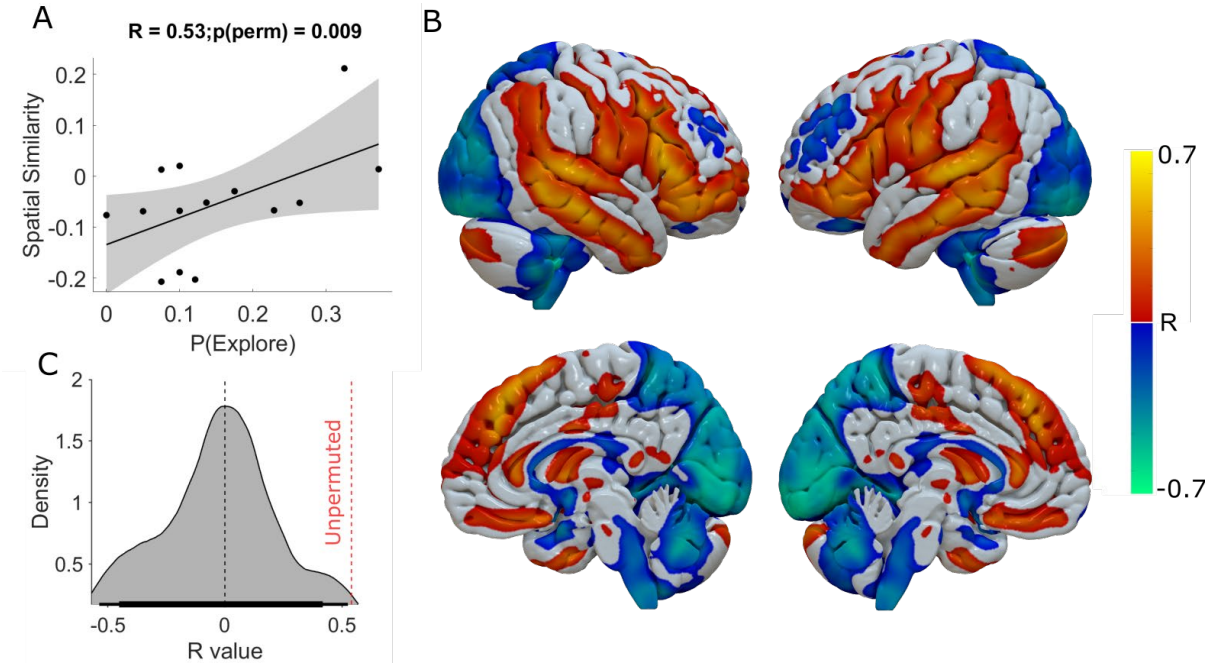

Figure 4 Functional connectivity of DBS-Induced exploration. The whole-brain voxel-wise R-Map demonstrates the optimal functional connectivity profile for DBS-induced enhacement of exploration. (**B)** Warm colours show voxels with functional connectivity to the DBS stimulation volumes associated with greater exploration. Cool colours indicate voxels with functional connectivity to the DBS stimulation volumes associated with lesser exploration. The more the individual functional connectivity profile matched the 'optimal' R-Map, the greater was the DBS-induced exploration (**A**) (R2 = 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)

Multiple effects of DBS on decision choices revealed by dissociable influence on RLDDM parameters

We performed posterior predictive checks to test whether the dual learning rate RLDDM model 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 5 (A & C). These represent the model's generative choices for fifty simulated "experiments" with the RLDDM model 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) overlap with the observed choices (solid line) for both conditions. Furthermore, because the RLDDM model 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 supported that the RLDDM model captured the salient features of the patient's 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.

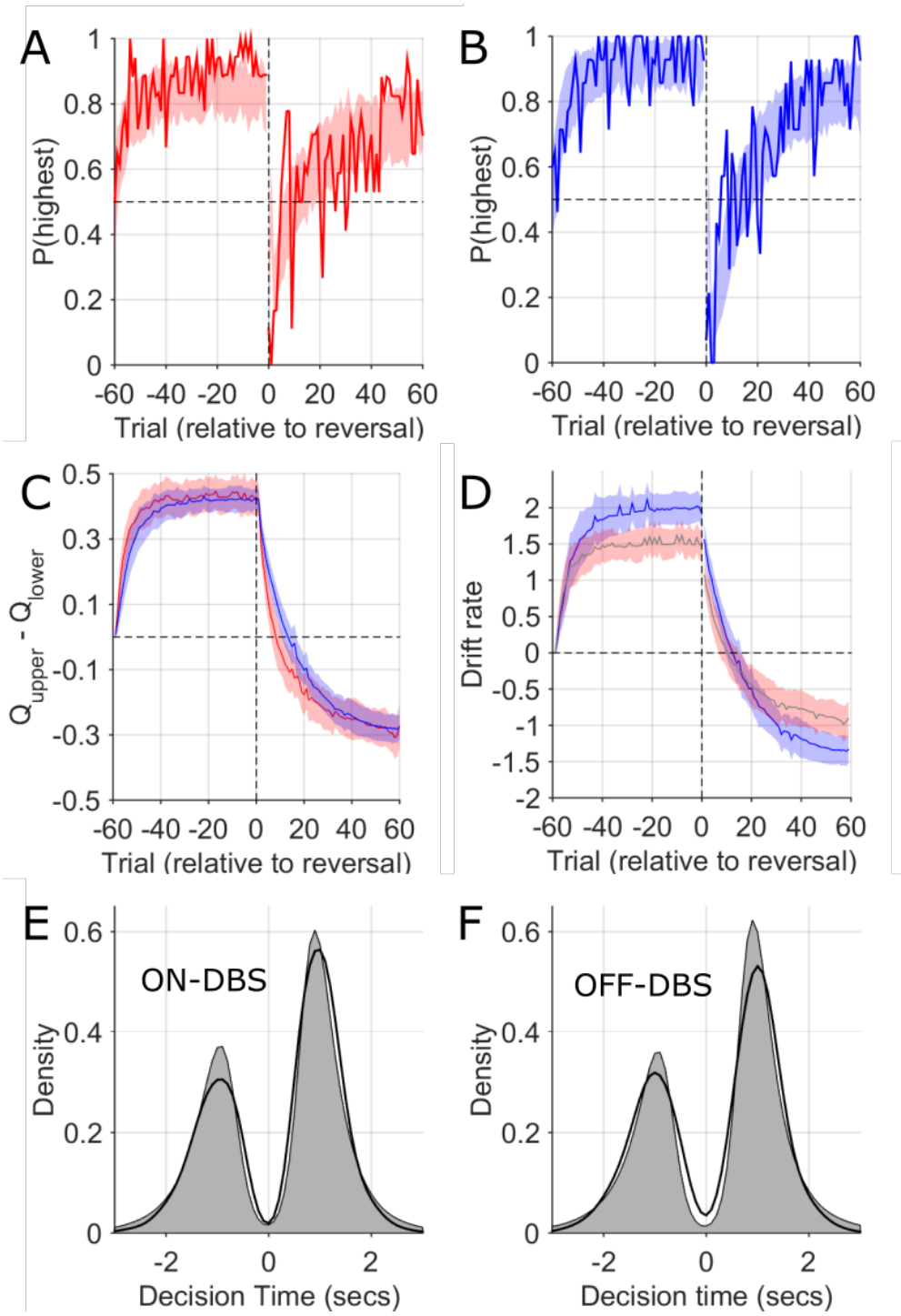

Figure 5 Observed and simulated RLDDM data. (A) The probability of choosing the highest value stimulus in the ON-DBS state is represented by the solid red line. Overlaid shaded area represents 95% confidence limits of stimulated choices of the RLDDM model (averaged across 50 simulated experiments). The same analysis is for the OFF-DBS state is plotted in (**B**). 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 conditions. (**C**) The difference in the expected value (Q) of the two choices in the RLDDM model (mean±95% confidence limits) represented by the solid blue (OFF-DBS) and red (ON-DBS) lines with the shaded overlay the confidence limits estimated from the synthetic (simulated) data. (**D**) The trial-to-trial variation in the mean drift rate, *v,* across the simulation in both the ON and OFF conditions demonstrates a reduction in the ON-DBS drift rate consisent with the group level effect of DBS on reducing the drift rate scaleing parameter, *m,* see Figure 6 and Table 1. Thicker back lines display the observed decision time distribution across all patients and sesssion in the ON-DBS **(E)** and OFF-DBS **(F)** conditions, with the simulated DT from the RLDDM model represented by the grey shaded regions. The decision time (DT) for choices which have

an initially lower value are are shown as negative. The reliablity of the RLDDM model in capturing the deicsion mechanisms in the task are supported by the overlap in observed and simulated DT distributions.

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 dependant variables of DBS state (ON or OFF). The non-decision time, *t*, was estimated for each subject but was assumed not to be influenced by DBS state. The mean estimated non-decision time was 0.57± 0.24 seconds. Group level parameters of the within-subject effects of DBS were used to assess how pallidal neuromodulation influenced exploratory choices (Figure 6, Table 2). We calculated Bayes Factors as measures of evidence for the existence of directional effects of DBS on the posterior distribution of the model parameters [42, 50]. Applying Jeffrey's evidence categories for Bayes Factors [51] the within subject's comparison demonstrated very strong evidence (BF = 37.9) for an increase in the positive learning rate, α_+, in the ON-DBS state. The results also indicated substantial evidence for a decrease in both the drift rate scaling parameter, *m*, (BF = 8.2) and boundary separation parameter, *a*, (BF = 5.6). There was no evidence in support of an effect of DBS on the negative learning rate α_-.

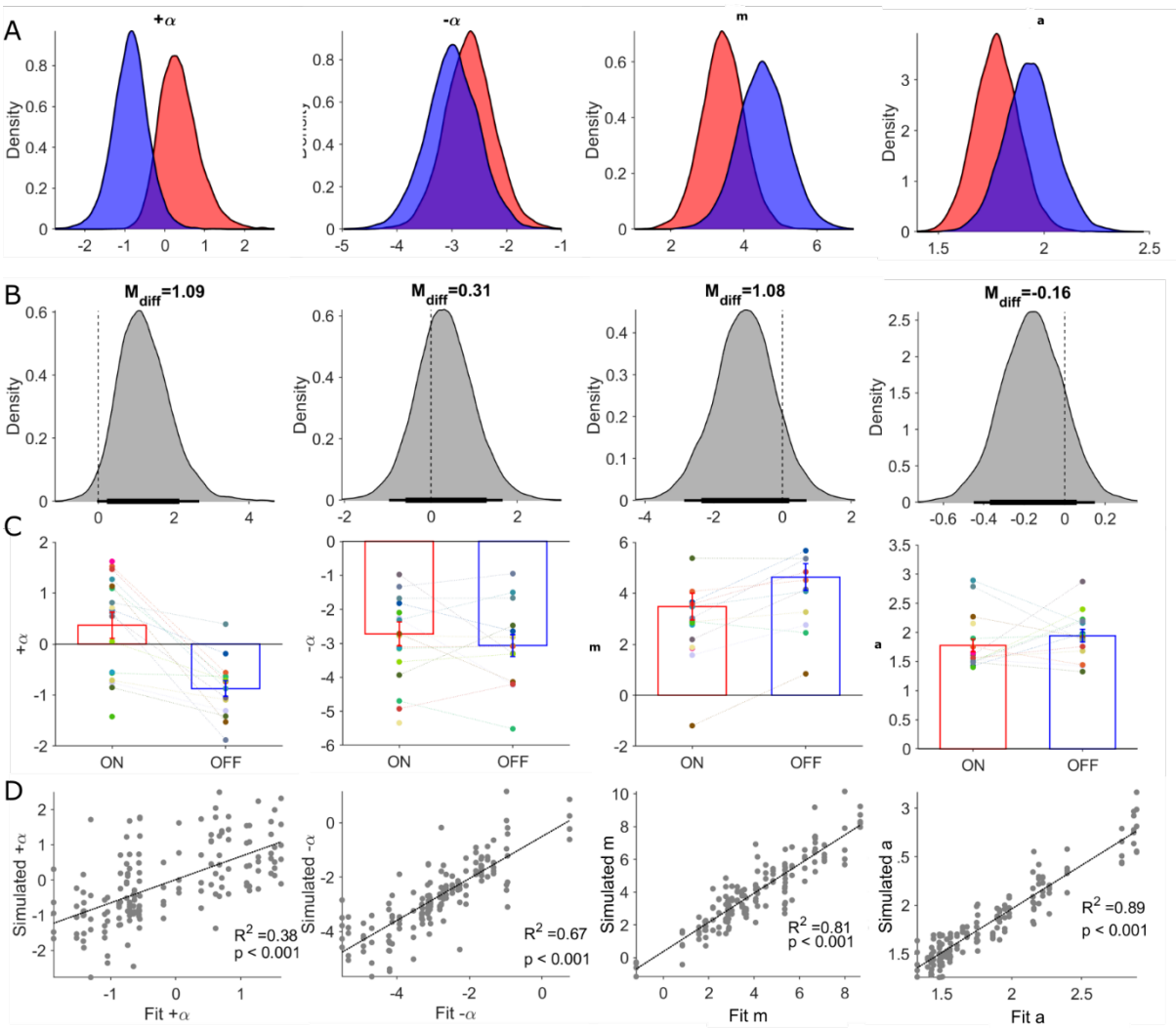

Figure 6 Posterior distributions of RLDDM parameters and parameter recovery. Columns in this figure display results for each of the estimated parameters in the RLDDM model from left to right including, positive learning rate (α_+), negative learning rate (α_-), drift rate scaling parameter (m), boundary 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 of and validation of applying the RLDDM model of decision making.

Table 1 Summary of posterior distributions

	ON-DBS			OFF-DBS			Contrast			
Parameter	***m***	**HDI**		***m***	**HDI**		***m***	**HDI**		**BF**
Boundary Separation	1.78	1.56	1.98	1.96	1.7	2.18	-0.16	-0.45	0.15	5.6
Drift rate Scaling	3.4	2.26	4.54	4.51	3.20	5.88	-1.08	-2.86	0.71	8.2
Learning rate +	0.22	-0.55	1.37	-0.85	-1.79	0.01	1.09	-0.03	2.68	37.9
Learning rate -	-2.65	-3.58	-1.8	-2.97	-3.97	-2.05	0.31	-0.97	1.66	2.16

Estimated means (m) for ON and OFF-DBS conditions as well as contrast for ON-OFF. HDI values are for 95% highest density Interval. Bayes Factors for directional effects are calculated as i/(1-i), where i is the integral of the posterior distribution.

To reconcile DBS changes in the model parameters with enhanced exploration we generated simulated choices and estimated the P(Explore) values across multiple combinations of the parameter space (Figure 7). As expected from the RLDDM choice mechanism, lower boundary separation values led to a higher proportion of exploratory choices driven by noisy perturbations reaching the lower value choices decision boundary (Figure 7B&C). Decreasing the drift rate scaling parameter (m) also drove increases in exploration (Figure 7 A&C). Mechanistically this is consistent with a weakened influence of the expected value (i.e. reduced exploitation) on the drift rate, so that choices are less biased by the differences in their value and more vulnerable to noise in the decision (diffusion) process. Both predicted effects of these simulations were seen in the DBS ON condition as indexed by the reduction in their mean posterior probability values for *m* and *a* parameters. In contrast, simulating P(Explore) values across a range of α_+ values revealed no significant influence other than a reduction in P(Explore) at the extreme lower values of the range tested. These simulations confirmed that the increased α_+ values in the ON-DBS were likely to be explained by a dissociable effect of DBS on value representation within the GPI that was independent of the enhanced exploratory choices in the ON state.

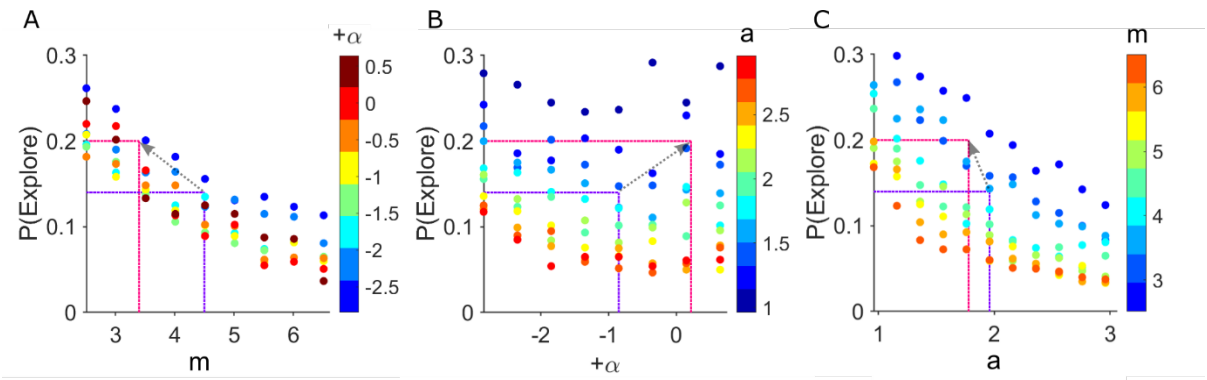

Figure 7 Mapping the parameter space interactions and effect on exploration in the RLDDM. (A) The interaction between the drift rate scaling parameter (m) and the positive learning rate across a range of parameter values. Colour scale represent the different values of positive learning rate. The experimentally observed increase in P(Explore) values across sessions is superimposed for ON-DBS (red line) and OFF-DBS (blue) conditions. The increase in exploration is driven by the reduction in boundary parameter and drift rate scaling (m) parameters in the ON-DBS state (**A**) and (**C**). In contrast, increased positive learning rate values (**B)** have no effect on exploration within the parameter range influenced by DBS.

Discussion

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

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

The exploration – exploitation "dilemma" is considered one of the fundamental challenges of adaptive control and behaviour [1, 2]. We took advantage of the opportunity leveraged by DBS to address the paucity of experimental evidence[19] relative to theoretical support for the BG role in explore-exploit choice arbitration[16-18].

Studies of patients and non-human primates with lesions of the GPI exhibit an inability to learn novel stimulus response contingencies and random choice behaviour which prevents the acquisition of choice preference or implicit learning of sequences [39, 52]. We did not observe levels of performance degradation in our patients that would be comparable to complete lesioning of pallidal output. Neurophysiological data support a net inhibitory effect on GPI neuron firing rates without complete firing suppression at stimulation intensities comparable to those used in our patients [53]. We suggest that the most likely neuromodulatory mechanism of DBS on enhanced exploration is a partial inhibition of GPI firing and that the behavioural consequences observed in our study are proportionate to the degree of electrically lesioning the basal ganglia's output [54-57].

Previous studies have confirmed that during acute changes in GPI-DBS (comparable to the ON-OFF conditions tested in our patients) a reduced inhibitory output induced by DBS in GPI leads to increased excitability of the motor cortex [58, 59]. Motor cortex activity correlates negatively with decision value during explore-exploit decisions [26]. This means that within a sequential sampling decision mechanism, increased motor cortex activity correlates with small differences in the value of the two choices and corresponding higher decision uncertainty and choice exploration. This was indeed the 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 [42]. 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 [60] propose that activity in the direct (striato-nigral) and indirect (striato-pallidal) BG pathways differentially encode the rate of evidence accumulation and amount of information needed to reach decision threshold. Our finding that GPI-DBS reduced both the drift rate scaling and boundary separation parameters, which govern these properties of the RLDDM model, would be in agreement with this idea. However, because the GPI sits anatomically at the point of convergence of both these 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 proximal to the GPI, such as the subthalamic nucleus (STN) could be used to delineate this mechanism further.

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 becomes habitual and stimulus response mapping is obtained [19]. These findings are consistent with physiological changes in the direct and indirect pathways during different stages of learning. Yin et. al.,[61] 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 [62] increasing its excitability and promoting exploitation. Recent data from Lee and Sabatini (2021) [63] 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 [64] and shifting [65].

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 [5].This mechanism could be supported by the observed decrease in the boundary separation parameter in the RLDDM model ON-DBS as this would increase the model's sensitivity to noisy perturbations in the process of reaching the decision bounds. Examining the effects of GPI-DBS in a non-stationary 'restless' bandit task could delineate the contribution of decision noise as these have models take noise into account of the decision process have been successfully fitted to such tasks[66] .

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

We found that the learning rate to rewarding outcomes was enhanced by DBS. This meant that the patient's choices were weighted more strongly by recent outcomes that were better than expected. Mapping the parameter space of the RLDDM model suggested this was unlikely to contribute to the

398 main behavioural shift towards exploration supporting a dissociable effect of DBS on GPI function.
399 Recordings of Local Field potentials (LFPs) from DBS electrodes in the GPI demonstrate that these
400 encode choice outcome in patients with impulsivity [67]. Reward Prediction Error (RPE) encoding in
401 LFP's in GPI have also been identified and are most marked during exploratory decisions and absent
402 when choices are exploitative [68]. DBS in our patients may therefore have interfered with this intrinsic
403 GPI reward encoding through blocking co-lateral dopaminergic signalling from the nigro-striatal
404 pathway, which traverses the GPI on route to the striatum [69]. Alternatively, polysynaptic
405 neuromodulation of the PPN (Pedunculopontine Nucleus) via the pallido-peduncular tract could explain
406 this effect as DBS of the PPN enhances reward learning but not loss avoidance [70].

407 We failed to find any evidence of DBS effect on learning related to negative outcomes, indexed by the
408 selective increase in the positive learning rate ON-DBS, with no effect on the negative learning rate.
409 This is surprising given that pallidal error-related activity proceeds the cortical error-related negativity
410 (ERN), which is, in turn, closely related to the negative prediction error signal [71, 72]. Although
411 differences in the task could partly explain this lack of effect, negative prediction error signalling may
412 be influenced by the GPI – lateral habenula pathway [73]. The relative sparseness of this projection,
413 compared to thalamic projections from GPI may have made this less amenable to neuromodulatory
414 blockade by DBS.

415 *Integrating the results in a network perspective*

416 The postero-lateral portion of the GPI targeted in the group of patients studied here is chosen in routine
417 clinical practice because of its position within the sensorimotor loop of the basal ganglia [74] to relieve
418 motor symptoms of dystonia. It is noteworthy that despite evidence for anatomically segmented regions
419 of the GPI (based on white matter connectivity [32]) stimulating the sensorimotor pallidum in this study
420 could alter decision-making and therefore modify "cognitive" function. At first glance, this result might
421 appear incongruous with its established role in motor control. However, our functional connectivity
422 analysis supports the conclusion that sensorimotor GPI function extends beyond that traditionally
423 attributed to it, on the basis of anatomical connectivity. The finding of a neuromodulatory influence of
424 sensorimotor pallidum on decision making is consistent with fMRI activation of pre-and post-central gyri,
425 during exploratory choices [24] and the encoding of decision value by motor cortex [75].

426 Furthermore, there is increasing evidence that DBS exerts influence on remote areas beyond the
427 stimulation site [48]. For instance, modulation of distributed networks through subthalamic DBS affects
428 motor learning and risk-taking behavior in Parkinson's disease patients [76, 77]. The exploration
429 enhancing effects of GPI-DBS in our study were predicted by functional connectivity to brain regions
430 whose neurons encode uncertainty [27] and predict behavioural switching[29, 30] – functions
431 synonymous with exploratory information enhancing decisions. Thus, the basal ganglia integrate
432 functionally segregated distributed networks that can be modulated by DBS.

433 *Clinical relevance - apathy.*

434 From a clinical perspective, these findings may be relevant to understanding the mechanisms which
435 worsen apathy in Parkinson's disease (PD) following DBS targeting the basal ganglia [78]. Patients

with PD-Apathy tend to preferentially choose high value outcomes over options with lower payouts[79]. In the DBS-ON 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 [80].

Study limitations

Patients with primary forms of dystonia who have not undergone DBS exhibit abnormal reinforcement learning [34, 81, 82]. Thus, although we cannot discount the possibility that the effect of DBS in this study is not related to interaction with a pathologically biased basal ganglia circuit, reversal-learning deficients in focal dystonia are most likely due to enhanced exploitation [83].This would make the behavioural effects of DBS in this study are more likey an underestimate of the influence of changes in pallidal excitability on decision-making.

The use of normative connectomes instead of patient specific data enhance the signal to noise ratio and image quality [48] 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 [84, 85]. 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 subcortical and cortical circuits to explore-exploit decision-making as well as their interaction, leading to insights into disordered decision-making in psychiatric and neurologic conditions [86].

Materials and Methods

Participants

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

see supplementary Table 1. In accordance with the declaration of Helsinki, participants gave written informed consent to participate in the study, which was approved by the local ethics committee (Charité – Universitätsmedizin Berlin, EA1/179/20).

Experimental Task

We used a modified version of a two-choice reward reversal-learning task based on Pessiglione et. al., (2006) [87]. The task was presented on a laptop screen using Psychtoolbox v3.0 [88] running on MATLAB (R2019, The MathWorks, Natick, MA, USA). The laptop was positioned in front of the patient so that responses could be made on the keyboard. Participants were instructed to try to win as many "vouchers" as possible. Printed screen text was in German. At the beginning of each trial a fixation cross presented at the centre of the screen indicated that a new trial had begun. A pair of fractal images were presented and subjects were expected to indicate their choice by a keyboard button press. The order of the fractals was randomly assigned to either the right or left of the fixation cross. In the event that they did not make a response within 3 seconds the fixation cross was re-presented and a new trial began. A choice was highlighted for 0.5 seconds by a red circle around the chosen fractal. The choice screen was followed by feedback screen for a further 1 second. Feedback consisted of either – "you win" or a neutral feedback condition with a screen with the words "nothing". Patients performed three sessions of 40 trials with 3-4 minute breaks between sessions to improve task compliance. The visual stimuli were associated with a fixed probability of rewarding outcome of 80:20%. Reward contingencies were reversed after 60 trial presentations (i.e. the image associated with a 80% reward probability in the first 60 trials was associated with 20% reward probability in the last 60 trials and vice-versa). A short training session of 10 trials with a novel pair of fractals was performed before formal testing began to familiarise the 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[44] ;

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

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 [42] 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 [89]. 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 [90]. For all estimates including those for parameter recovery the Gelman Rubin values were all less than 1.1 indicating successful convergence.

Posterior distributions in the differences in the parameter estimates were estimated by subtracting the posterior estimates for each parameter in the ON and OFF-DBS states. The posterior means and highest density intervals (HDI) were calculated using the 'bayestestR' and 'tidybays' packages in R version 4.1.1 [91, 92]. Bayes Factors for directional effects were calculated as *i*/(1-*i*) where *i* is the integral of the posterior distribution from 0 to +∞ for values as previously reported [42, 93].

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

Lead localization

DBS Electrodes were localized for 14/19 of the patients whom both testing data was available ON and OFF-DBS using the software "Lead-DBS" V2.5 [43] www.lead-dbs.org as described [76, 94] and mapped in MNI template space. Stimulation volumes were modeled based on SimBio-FieldTrip pipeline incorporated in Lead DBS [34]. Stimulation volume modeling was informed by DBS parameters used to control patients' symptoms during behavioral task (Supplementary Table 1). Each patient's

547 stimulation volume were then used as seed regions estimate voxel-wise functional connectivity profile
548 maps using an openly available group connectome derived from resting state functional connectivity
549 MRI images of 1000 neurologically healthy volunteers[95]. Initially, whole-brain functional connectivity
550 maps from stimulation volumes were created for each patient. Then, voxels whose connectivity was
551 correlated with DBS-induced changes in the probability of exploring the lower value choice "P(Explore)"
552 were visualized as R-maps [48, 76, 94]. The DBS induced change in P(Explore) was defined as the
553 maximum increase in P(Explore) ON-DBS minus P(Explore) OFF-DBS in one of the three sessions.
554 The similarity between the "ideal" connectivity profile of this R-map (Figure 4B) and the connectivity
555 profile of each individual patient was mathematically assessed by calculating spatial correlation co-
556 efficients' across all brain voxels [48, 96]. Finally, the predictive potential of the DBS connectivity maps
557 for the increase in exploration was estimated by correlating the individuals increase in P(Explore) with
558 DBS to the spatial correlation in each patient (Figure 4A). A permutation distribution was constructed
559 from the same analysis performed 1000 times and the correlation coefficient re-estimated with the
560 P(Explore) values for each subject randomly re-ordered.

561 **Behavioural data analysis**

562 Prior to statistical analyses, trials without responses (errors of omission) were excluded (ON-DBS mean
563 1.06±2.26 (range 0-8) trials, OFF-DBS 1.78±2.5 (range 0 - 7 trials). Accordingly, DT and task
564 performance (number of rewards won in a session) were analyzed using ANOVA with the MATLAB
565 function *anovan*. In case of significant interactions, post-hoc tests were conducted using paired samples
566 t-tests. Normality assumptions were tested using Kolmogorov Smirnov tests (all P > 0.05). All results
567 are reported as mean values ±S.E.M.

568 **Acknowledgements**
569
570 TPG was funded by a NRS Fellowship from the Chief Scientist Office (Scotland). Additional funding
571 for this research was provided by the University of Dundee/NHS Tayside Movement disorders
572 research Endowment Fund.

573
574 **Competing interests**
575
576 TPG has no competing interests.
577 ALAM has no competing interests.
578 JDS has no competing interests.
579 OG has no competing interests.
580 BA has no competing interests.
581 AAK has received from honoraria from Boston Scientific, Medtronic and Teva.

582
583 **References**

584 1. Mehlhorn, K., et al., *Unpacking the exploration–exploitation tradeoff: A synthesis of human*
585 *and animal literatures.* Decision, 2015. **2**(3): p. 191.
586 2. Cohen, J.D., S.M. McClure, and A.J. Yu, *Should I stay or should I go? How the human brain*
587 *manages the trade-off between exploitation and exploration.* Philosophical Transactions of
588 the Royal Society B: Biological Sciences, 2007. **362**(1481): p. 933-942.

3. Herrnstein, R.J., *On the law of effect 1.* Journal of the experimental analysis of behavior, 1970. **13**(2): p. 243-266.
4. Daw, N.D., et al., *Cortical substrates for exploratory decisions in humans.* Nature, 2006. **441**(7095): p. 876-9.
5. 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.
6. Costa, V.D., et al., *Reversal learning and dopamine: a bayesian perspective.* Journal of Neuroscience, 2015. **35**(6): p. 2407-2416.
7. Izquierdo, A., et al., *The neural basis of reversal learning: an updated perspective.* Neuroscience, 2017. **345**: p. 12-26.
8. Sutton, R.S. and A.G. Barto, *Reinforcement learning: An introduction*. 2018: MIT press.
9. 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.
10. Findling, C., et al., *Computational noise in reward-guided learning drives behavioral variability in volatile environments.* Nat Neurosci, 2019. **22**(12): p. 2066-2077.
11. 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.
12. Costa, V.D., et al., *Dopamine modulates novelty seeking behavior during decision making.* Behavioral neuroscience, 2014. **128**(5): p. 556.
13. 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.
14. 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.
15. 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.
16. Gilbertson, T. and D. Steele, *Tonic dopamine, uncertainty and basal ganglia action selection.* Neuroscience, 2021. **466**: p. 109-124.
17. 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.
18. 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.
19. Sheth, S.A., et al., *Basal ganglia neurons dynamically facilitate exploration during associative learning.* Journal of Neuroscience, 2011. **31**(13): p. 4878-4885.
20. 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.
21. 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.
22. 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.
23. Ghahremani, D.G., et al., *Neural components underlying behavioral flexibility in human reversal learning.* Cerebral cortex, 2010. **20**(8): p. 1843-1852.
24. Chakroun, K., et al., *Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making.* Elife, 2020. **9**.
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. 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.
35. Romano, R., et al., *Impaired cognitive functions in adult-onset primary cranial cervical dystonia.* Parkinsonism & related disorders, 2014. **20**(2): p. 162-165.
36. 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.
37. Lafreniere-Roula, M., et al., *High-frequency microstimulation in human globus pallidus and substantia nigra.* Experimental brain research, 2010. **205**(2): p. 251-261.
38. Sheth, S.A., et al., *Basal ganglia neurons dynamically facilitate exploration during associative learning.* J Neurosci, 2011. **31**(13): p. 4878-85.
39. 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.
40. 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.
41. Daw, N.D., et al., *Model-based influences on humans' choices and striatal prediction errors.* Neuron, 2011. **69**(6): p. 1204-1215.
42. 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.
43. Horn, A., et al., *Lead-DBS v2: Towards a comprehensive pipeline for deep brain stimulation imaging.* Neuroimage, 2019. **184**: p. 293-316.
44. Rescorla, R.A., *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.* Current research and theory, 1972: p. 64-99.
45. Ratcliff, R., *A theory of memory retrieval.* Psychological review, 1978. **85**(2): p. 59.
46. Miletić, S., et al., *A new model of decision processing in instrumental learning tasks.* Elife, 2021. **10**: p. e63055.
47. 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.
48. Horn, A., et al., *Connectivity predicts deep brain stimulation outcome in P arkinson disease.* Annals of neurology, 2017. **82**(1): p. 67-78.
49. 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.
50. Stefan, A.M., et al., *A tutorial on Bayes Factor Design Analysis using an informed prior.* Behavior Research Methods, 2019. **51**(3): p. 1042-1058.
51. Jeffreys, H., *The theory of probability*. 1998: OUP Oxford.
52. 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.

53. 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.
54. 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.
55. Dostrovsky, J., et al., *Microstimulation-induced inhibition of neuronal firing in human globus pallidus.* Journal of neurophysiology, 2000. **84**(1): p. 570-574.
56. 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.
57. 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.
58. 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.
59. Kühn, A., et al., *Modulation of motor cortex excitability by pallidal stimulation in patients with severe dystonia.* Neurology, 2003. **60**(5): p. 768-774.
60. 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.
61. 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.
62. Kravitz, A.V., et al., *Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry.* Nature, 2010. **466**(7306): p. 622-626.
63. Lee, J. and B.L. Sabatini, *Striatal indirect pathway mediates exploration via collicular competition.* Nature, 2021. **599**(7886): p. 645-649.
64. Tecuapetla, F., et al., *Complementary contributions of striatal projection pathways to action initiation and execution.* Cell, 2016. **166**(3): p. 703-715.
65. 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.
66. Findling, C., et al., *Computational noise in reward-guided learning drives behavioral variability in volatile environments.* Nature neuroscience, 2019. **22**(12): p. 2066-2077.
67. Eisinger, R.S., et al., *Distinct Roles of the Human Subthalamic Nucleus and Dorsal Pallidum in Parkinson's Disease Impulsivity.* Biological Psychiatry, 2021.
68. 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.
69. Lindvall, O. and A. Björklund, *Dopaminergic innervation of the globus pallidus by collaterals from the nigrostriatal pathway.* Brain research, 1979. **172**(1): p. 169-173.
70. Skvortsova, V., et al., *A Causal Role for the Pedunculopontine Nucleus in Human Instrumental Learning.* Current Biology, 2021. **31**(5): p. 943-954. e5.
71. 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.
72. Holroyd, C.B., et al., *Errors in reward prediction are reflected in the event-related brain potential.* Neuroreport, 2003. **14**(18): p. 2481-2484.
73. Hong, S. and O. Hikosaka, *The globus pallidus sends reward-related signals to the lateral habenula.* Neuron, 2008. **60**(4): p. 720-729.
74. 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.

75. 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.
76. 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.
77. Irmen, F., et al., *Sensorimotor subthalamic stimulation restores risk-reward trade-off in Parkinson's disease.* Movement Disorders, 2019. **34**(3): p. 366-376.
78. 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.
79. 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.
80. 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.
81. Arkadir, D., et al., *DYT1 dystonia increases risk taking in humans.* Elife, 2016. **5**: p. e14155.
82. 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.
83. 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.
84. Gershman, S.J., *Deconstructing the human algorithms for exploration.* Cognition, 2018. **173**: p. 34-42.
85. Schulz, E. and S.J. Gershman, *The algorithmic architecture of exploration in the human brain.* Current opinion in neurobiology, 2019. **55**: p. 7-14.
86. 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.
87. Pessiglione, M., et al., *Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans.* Nature, 2006. **442**(7106): p. 1042-1045.
88. Brainard, D.H. and S. Vision, *The psychophysics toolbox.* Spatial vision, 1997. **10**(4): p. 433-436.
89. 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.
90. 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.
91. Kay, M., *tidybayes: Tidy data and geoms for Bayesian models.* R package version, 2019. **1**(0).
92. 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.
93. Wiehler, A., K. Chakroun, and J. Peters, *Attenuated directed exploration during reinforcement learning in gambling disorder.* Journal of Neuroscience, 2021. **41**(11): p. 2512-2522.
94. Neumann, W.-J., et al., *Functional segregation of basal ganglia pathways in Parkinson's disease.* Brain, 2018. **141**(9): p. 2655-2669.
95. Holmes, A.J., et al., *Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures.* Sci Data, 2015. **2**: p. 150031.
96. Al-Fatly, B., et al., *Connectivity profile of thalamic deep brain stimulation to effectively treat essential tremor.* Brain, 2019. **142**(10): p. 3086-3098.

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.

853