Dopamine depletion in Parkinson's increases directed but not random exploration

Björn Meder[1*], Martha Sterf[2], Charley M. Wu[3,4,5], Matthias Guggenmos[1]

[1*]Institute for Mind, Brain, and Behavior, Health and Medical University, Olympischer Weg 1, 14471 Potsdam, Germany.
[2]Medical School Berlin, Rüdesheimer Str. 50, 14197 Berlin, Germany.
[3]Human and Machine Cognition Lab, University of Tübingen, Maria-von-Linden-Str. 6, 72074 Tübingen, Germany.
[4]Institute of Psychology, Technical University of Darmstadt, Karolinenplatz 5, 64289 Darmstadt, Germany.
[5] Hessian.AI, Landwehrstraße 50A, 64293 Darmstadt, Germany.

*Corresponding author(s). E-mail(s): bjoern.meder@hmu-potsdam.de;
Contributing authors: martha.sterf@student.medicalschool-berlin.de;
charley.wu@tu-darmstadt.de; matthias.guggenmos@hmu-potsdam.de;

Abstract

We investigated how patients with Parkinson's disease (PD) manage the explore–exploit trade-off in a structured reward-learning task. Patients were tested either on (N=34) or off (N=34) dopaminergic medication (levodopa), with age-matched polyneuropathy patients serving as controls (N=35). Behaviorally, patients off medication showed marked learning and decision-making deficits, characterized by overexploration and insufficient exploitation. To clarify the mechanisms underlying these impairments, we applied a computational model that combines similarity-based generalization with both random and uncertainty-directed exploration. The modeling results showed that impairments in patients off medication resulted from reduced generalization and increased uncertainty-directed exploration, but not greater random exploration. In contrast, exploration and generalization in patients on medication were comparable to the control group. Our findings highlight how dopamine depletion in PD impacts reward learning under uncertainty, suggesting a key role of dopamine in exploration and generalization.

Keywords: explore-exploit trade-off, Parkinson's disease, levodopa, dopamine, uncertainty-directed exploration, random exploration, generalization, reinforcement learning, multi-armed bandit

Parkinson's disease (PD) involves degeneration of the dopaminergic system, which is central not only for motor control but also for learning, decision making, and exploratory behavior [1–6]. The dopamine deficit in PD impairs several aspects of cognitive function, including the ability to learn from feedback and adapt to changes in reward contingencies [7–9]. While simpler associative learning mechanisms can be preserved, patients show marked deficits in tasks that require building and using an internal model of the environment [10]. Dopaminergic medication in the form of levodopa (L-Dopa) can alleviate these impairments to some degree, especially when learning from reward feedback [7, 11–13].

Critically, PD also interferes with the ability to navigate the exploration–exploitation trade-off [14–16], which requires balancing exploring promising novel options and exploiting known high-reward options. Whether deciding between ordering your favorite dish or trying something new, or choosing a new travel destination over a familiar favorite—the fundamental challenge is to choose between what is known to work and discovering something even better [17–19]. Investigating how PD influences the exploration–exploitation trade-off is therefore important from both a clinical and theoretical perspective, as it links dopaminergic dysfunction to central mechanisms of learning and decision making.

An essential paradigm for investigating explore–exploit problems are *bandit tasks*, in which participants learn about choice options with uncertain outcomes through trial and error [20–23]. In each trial, learners must choose between the available options, balancing the exploration of new options with the exploitation of known options. While many tasks in the literature have independent options where the outcome of one decision does not provide information about other options [24–28], natural environments are often characterized by structural regularities [29–31]. For instance, spatially proximate areas often yield similar foraging outcomes [32], and if two dishes share similar ingredients, trying one helps predict the taste of the other [33]. Such latent regularities can be utilized to generalize beyond direct experiences and adapt exploratory decisions to environmental structure [34].

Exploratory behavior can be decomposed into two distinct mechanisms: *random exploration* that results from inherent decision noise, and *uncertainty-directed exploration*, which is driven by information value [29, 35–37]. As adaptive decision-making under uncertainty requires balancing both forms of exploration, a central question is how exactly the dopaminergic system shapes exploratory behavior. For instance, one previous study found that levodopa selectively reduced uncertainty-directed exploration in healthy participants, while random exploration remained largely unaffected [26]. Moreover, the influence of valence on information acquisition can be reduced under levodopa [38]. These findings suggest that dopamine modulates the valuation processes that govern exploratory decisions. However, the precise role of dopamine in exploratory behavior is still poorly understood.

Goals and scope

It is well established that patients with PD exhibit pronounced deficits in reinforcement learning, typically attributed to dopaminergic degeneration[7–13]. To better understand how dopamine and medication specifically shape exploratory behavior in PD, we used a behavioral paradigm that captures two critical features of real-life exploration: the abundance of choice options and the presence of hidden structure that can be leveraged to search efficiently.

To this aim, we tested three groups: PD patients off levodopa medication, PD patients on levodopa medication, and an unmedicated control group with polyneuropathies (peripheral nerve damage). By combining behavioral and computational analyses we isolate deficits specific to dopamine depletion in PD, evaluate the restorative effects of levodopa medication, and assess the extent to which medicated patients resemble controls.

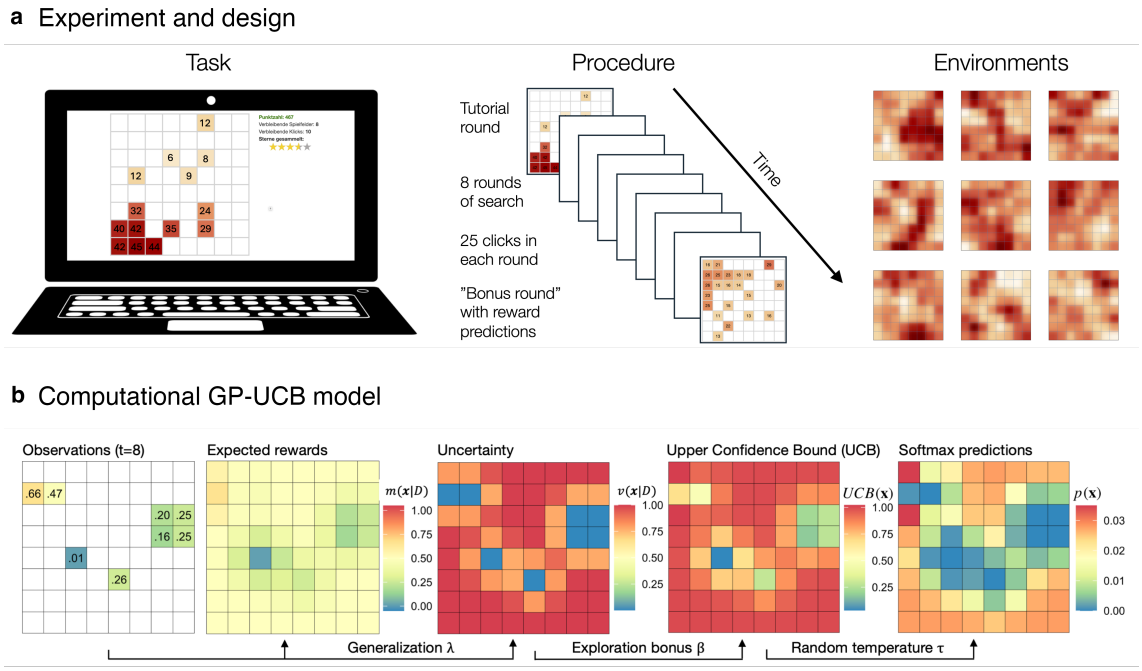

Fig. 1 Spatially-correlated multi-armed bandit task and computational model. **a.** Screenshot from the experiment where participants selected tiles (options) on a 8×8 grid to accumulate rewards. Each participant completed 10 rounds, where in each round a new environments was presented. The first round was a tutorial round and the last round was a bonus round where participants made predictions for 5 randomly selected tiles after making 15 explore-exploit choices. In each round, a new environment was used, drawn randomly from a set of 40 environments with the same level of spatial correlation. Darker shades of red indicate higher rewards. **b.** Task behavior was modeled using a Gaussian Process (GP) with Upper Confidence Bound (UCB) sampling. The GP-UCB model combines similarity-based generalization with mechanisms for uncertainty-directed and random exploration. First, the observed data are used to fit a Gaussian Process model, which estimates the expected rewards $m(\mathbf{x})$ and associated uncertainty $v(\mathbf{x})$ for each option (tile), conditional on the observed data. The amount of generalization is determined by parameter λ, the length-scale of the RBF kernel, which in the Gaussian Process determines how quickly reward correlations decay with spatial distance. Next, options are valued using Upper Confidence Bound (UCB) sampling, which inflates reward expectations with an uncertainty bonus β, representing the extent of uncertainty-directed exploration. Finally, UCB values are passed through a softmax choice rule, with the temperature parameter τ reflecting the degree of random exploration. Panel b is reproduced from [37] under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Results

We compared PD patients on and off medication and age-matched controls in their ability to perform a spatial reward learning task (Table 1). All PD patients (N=68) regularly received dopaminergic medication (levodopa) for symptomatic treatment and were randomly assigned to be tested either after taking their regular dose (PD+; N=34) or in a state of dopaminergic depletion shortly before their next dose (PD−; N=34). The control group consisted of age-matched patients with polyneuropathies, i.e. disorders or damage of the peripheral nerves without involvement of the central nervous system (N=35).

All three groups were of interest to our investigation. The comparison between PD− patients off medication and the control group indicates deficits specific to PD in a state of (natural) dopamine depletion. The comparison between PD patients off (PD−) and on medication (PD+) shows to what degree levodopa improves performance within PD. Finally, the comparison between PD patients

Table 1 Descriptive characteristics. Shown are means (SD); p-values pertain to χ^2 test for gender and one-way ANOVAs for the other variables.

	PD−	PD+	Control	p
N	34	34	35	
Gender (% Female)	15 (44%)	18 (53%)	21 (60%)	0.4
Age (years)	67.0 (6.5)	64.3 (6.2)	65.5 (7.0)	0.2
Depression (BDI-II)	8.6 (4.1)	8.6 (3.6)	8.5 (3.4)	>0.9
Cognitive functioning (MMSE)	28.8 (0.8)	29.1 (0.8)	28.8 (0.9)	0.3
PD severity (HY)	1.9 (0.7)	1.9 (0.6)		0.9
Levodopa EDD (mg/day)	809.0 (230.8)	894.0 (441.7)		0.6
Time since medication (min)	248.5 (32.7)	104.4 (59.2)		<0.001

Note. BDI-II = Beck Depression Inventory II [39]; HY = Hoehn-Yahr scale [40]; MMSE = Mini-Mental State Examination [41]; Levodopa EDD = Levodopa equivalent daily dose, calculated according to [42], for patients who consented to the access of their medical records (14 PD− and 12 PD+ patients).

on medication (PD+) and controls clarifies to what degree compensatory effects of levodopa align PD behavior to the control group.

Behavioral Results

Participants accumulated rewards by selecting tiles (options) on a 8×8 grid where the reward for each tile was drawn from a Gaussian distribution (Fig. 1). The spatial correlation between tiles could be used to optimize search behavior and efficiently navigate the exploration-exploitation trade-off. We analyzed behavior in terms of performance, temporal dynamics of exploration behavior, and spatial trajectories of search, based on eight rounds of 25 trials per participant.

Impaired reward learning in PD patients off medication

Hierarchical regression analyses revealed an effect of group on obtained rewards, but no effects of game round, PD severity, depressive symptoms, or cognitive functioning (SI; Tables S1 - S3). These variables were therefore not considered in subsequent analyses.

Fig. 2a shows the average reward across all trials. PD+ patients on medication achieved substantially higher rewards than PD− patients off medication ($t(66) = 5.9$, $p < .001$, $d = 1.4$, $BF > 100$), indicating a strong beneficial effect of levodopa. Controls achieved slightly higher rewards than PD+ ($t(67) = 2.6$, $p = .01$, $d = 0.6$, $BF = 4.5$) and were much better than PD− patients ($t(67) = 7.5$, $p < .001$, $d = 1.8$, $BF > 100$).

The stark performance disadvantage for PD− gradually developed across trials (Fig. 2b). While controls and PD+ showed relatively steep learning curves, the learning slope of PD− patients was barely positive, suggesting a substantial deficit in managing the exploration-exploitation trade-off.

Levodopa mitigates deficits in exploration and exploitation

We next analyzed the trade-off between exploration and exploitation. Fig. 2c shows that differences in reward accumulation are driven by learners' ability to adequately balance exploration (selecting novel options) and exploitation (re-clicking tiles). PD+ patients exploited substantially more than PD− patients (3% vs. 16%, $t(66) = 5.0$, $p < .001$, $d = 1.2$, $BF > 100$). Controls exploited even more than PD+ patients, although this effect was weaker (16% vs. 26%, $t(67) = 2.6$, $p = .01$, $d = 0.6$, $BF = 4.3$).

Mirroring the reward learning curves, controls and PD+ patients shifted more efficiently from exploring novel options to exploiting known options over time (Fig. 2d). Controls began exploiting earlier in the round and exhibited a stronger overall tendency toward exploitation compared to PD+ patients, explaining their relative performance advantage. In stark contrast, PD− patients

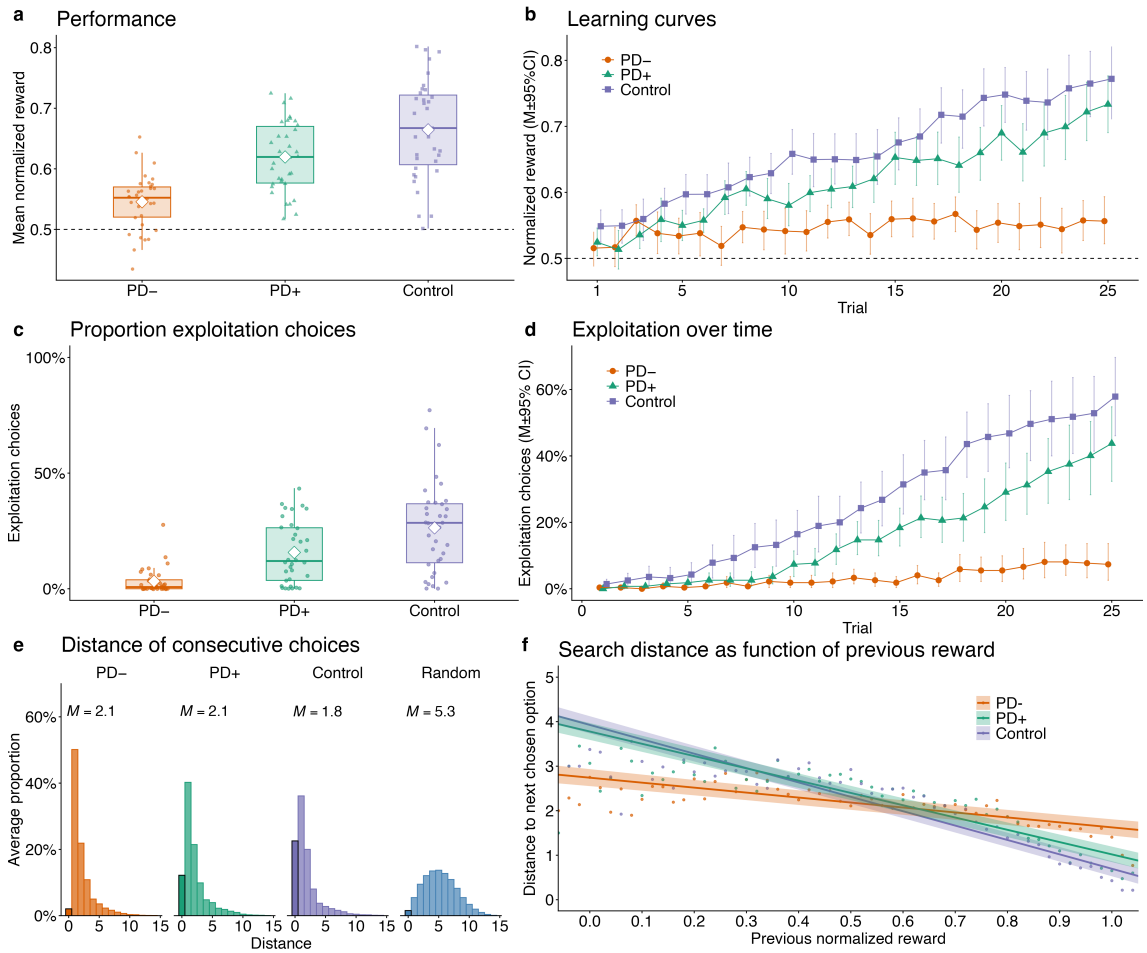

Fig. 2 Behavioral Results. **a.** Obtained rewards by group. Each dot is one participants' mean reward across all rounds and trials. Across all figures, box plots indicate the median (horizontal bar) the interquartile range (box) and the mean (white diamond). Whiskers extend from the box to 1.5 × interquartile range. The black and dotted horizontal line indicates random performance. **b.** Learning curves, showing obtained mean reward for each trial, averaged across rounds. Error bars for learning curves indicated 95% CI. **c.**. Mean proportions of exploitation decisions, aggregated over trials and rounds. Each dot is one participant. An exploitative choice is defined as as re-clicking a previously revealed tile. **d.** Mean proportion of exploitation decisions per trial, averaged over rounds. **e.** Manhattan distance among consecutive clicks. A repeat click has a distance of zero (marked by a black outline); clicks on neighboring tiles have a distance of 1, and distances >1 correspond to clicks further away from the previous click. The right-most panel shows the distribution of distances for a random learner that selects in each trial with uniform probability among all options. **f.** Predictions of a Bayesian hierarchical regression with search distance as a function of reward on the previous trial. Dots are the empirical mean distances for each reward value, averaged over participants, trials, and rounds.

predominantly engaged in exploration and showed only a weak tendency toward exploitation as the search horizon approached its end.

Consistent with these observations, nearly all controls (94%) and PD+ patients (88%) made at least one exploit choice, whereas only 65% of PD− patients did. Among participants who engaged in each choice type, PD+ patients obtained higher rewards than PD− patients during both exploration ($t(50) = 2.5$, $p = .016$, $d = 0.7$, $BF = 3.3$) and exploitation ($t(50) = 3.6$, $p < .001$, $d = 1.0$, $BF = 38$). Thus, when they exploited and explored, PD+ patients on medication showed higher efficiency in both behaviors. Similarly, controls earned higher rewards than PD− patients during exploration ($t(53) = 3.0$, $p = .004$, $d = 0.8$, $BF = 10$) and exploitation ($t(53) = 3.9$,

$p < .001$, $d = 1.1$, $BF = 83$). Controls and PD+ patients did not differ (both $p > .27$), indicating that the efficiency of exploration and exploitation was largely restored by levodopa medication.

Levodopa restores reward sensitivity during exploration

We analyzed the spatial trajectory of individuals' search processes by quantifying step-by-step distances and testing how obtained rewards shaped subsequent choices. Fig. 2e shows the distribution of distances among consecutive clicks, illustrating how participants explored the space. In all groups, the most frequent choice was to click on a directly neighboring tile (distance=1), indicating a localized search strategy. This locality bias was strongest in PD− patients off medication, who most frequently selected directly neighboring tiles. Across all choices, controls had lower search distances than both patients off medication ($t(67) = -2.6$, $p = .011$, $d = 0.6$, $BF = 4.2$) and on medication ($t(67) = -2.3$, $p = .023$, $d = 0.6$, $BF = 2.4$). The mean distance of the two PD groups did not differ ($p > .7$), but the distributions indicate different underlying strategies: fewer repeats (distance 0) but more near choices (distance 1) in PD− patients, balancing out against PD+ patients who made more exploit choices but clicked less often on directly neighboring tiles. Simulations of a random learner show a very different pattern compared to human behavior (Fig. 2e, right panel), indicating that all groups deviated substantially from purely random choice behavior.

To better understand when participants chose to search locally versus farther away, we next examined how search distances depended on the reward obtained from the previous choice (Fig. 2f). If a large reward was obtained, learners should search more locally; if a low reward was obtained, learners should search farther away. This behavior was observed in both controls and PD patients on medication, showing how they leveraged the structure of the environment to guide their search process. In contrast, the exploration strategy for PD patients off medication was largely insensitive to reward magnitude, indicating a deficit in goal-directed exploration based on the structural regularities of the environment.

The Gaussian Process Upper Confidence Bound (GP-UCB) model captures key behavioral aspects of exploration and generalization

The behavioral analyses showed that individuals in a dopamine-depleted state exhibit a severe deficit in balancing exploration and exploitation. By contrast, the behavior of patients on medication was markedly improved and largely resembled controls. The increased exploration in patients off medication could reflect more random choice behavior, an increased emphasis on uncertainty-directed exploration, or a deficit in utilizing the structural regularities of the grid (i.e., impaired generalization).

To disentangle these mechanisms, we used the *Gaussian Process Upper Confidence Bound (GP-UCB) model* (Fig. 1b; see Methods for formal specification). The model integrates similarity-based generalization with two distinct exploration mechanisms: *uncertainty-directed exploration*, which seeks to reduce uncertainty about rewards, and *random exploration*, which adds stochastic noise without being directed towards a particular goal [29, 34]. These processes are captured by three key parameters: the generalization parameter λ (Eq. 2), which determines how strongly rewards are generalized across options; the uncertainty bonus β (Eq. 6), which governs the degree of uncertainty-directed exploration by determining the value given to uncertainty; and the temperature parameter τ (Eq. 7) which captures random exploration.

In previous studies using the same experimental paradigm, the GP-UCB model provided the best account of exploratory behavior in healthy participants [29, 35–37, 43, 44]. Importantly, by decomposing exploration into generalization (λ), uncertainty-driven exploration (β), and random exploration (τ), the model allows us to identify which mechanisms are altered by PD and medication. . In doing so, it directly connects to prior findings that levodopa impairs discrimination learning while sparing generalization in PD [45], that levodopa reduces directed exploration in healthy participants [26], and that PD disrupts the overall exploration-exploitation balance [14–16].

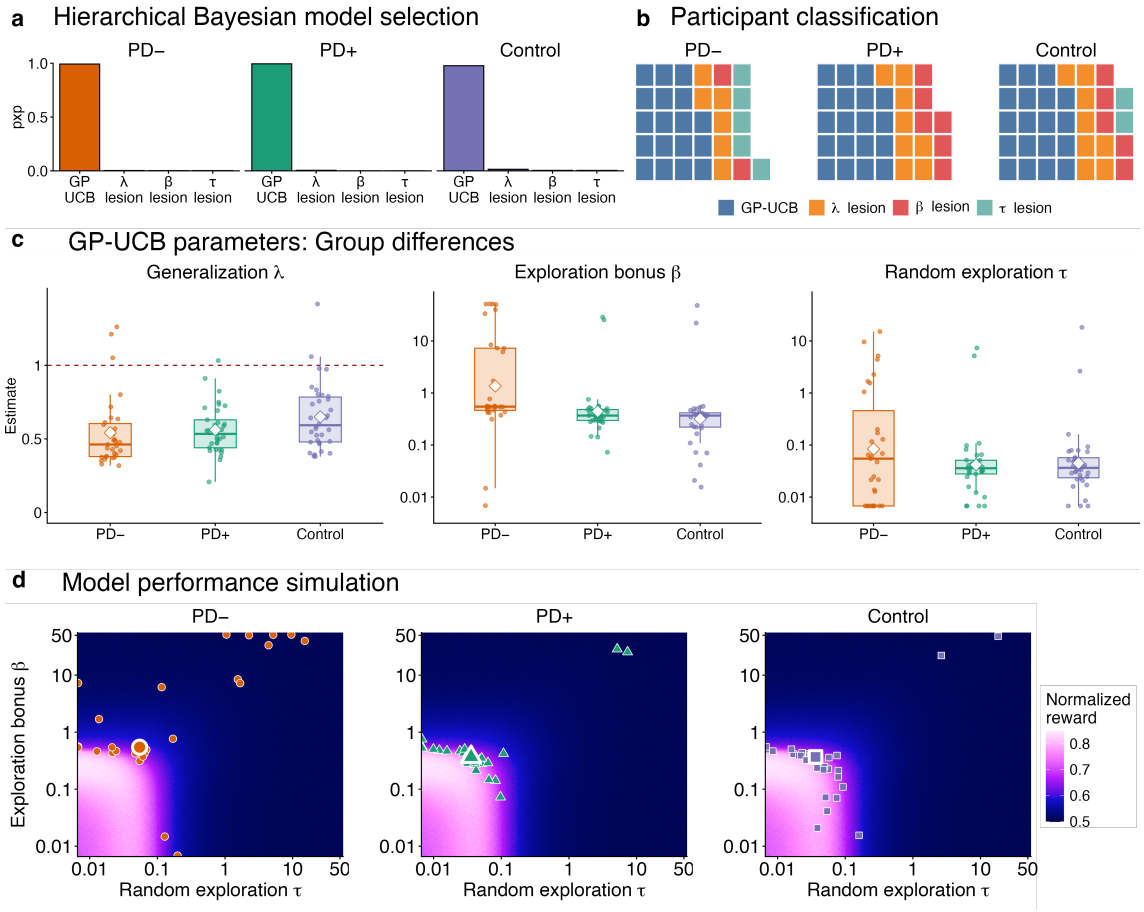

Fig. 3 Computational modeling results. **a)** Hierarchical Bayesian model comparison based on the protected exceedance probability (pxp) estimating which model is more frequent in the population. In each group, the GP-UCB model was the most likely model. **b)** Each participant was assigned to the model that best described their behavior (highest R^2 or, equivalently, lowest cross-validated log loss). In each group, the GP-UCB model accounted for the highest proportion of learners. **c)** Estimates for the generalization parameter λ, the uncertainty-directed exploration parameter β, and the random exploration parameter τ. The red dashed line in the λ plot denotes the true amount of spatial correlation ($\lambda = 1$). **d)** Performance landscape of the GP-UCB model across different amounts of random exploration τ and uncertainty-directed exploration β. The amount of generalization was fixed to $\lambda = 1$, the true amount of correlation in the environment, and β and τ were varied, with 1000 simulated learners per parameter combination. Each dot shows one participant; the larger markers indicate the group median.

All components of the GP-UCB model are critical to explain behavior

To assess the contribution of each model component (generalization, uncertainty-directed exploration, and random exploration), we compared the predictive accuracy of the GP-UCB model to variants where we lesion away each component (Methods).

The λ *lesion model* removes the ability to learn about the spatial regularities on the grid and thus to generalize. Instead, it is assumed that all options are learned independently (i.e., Bayesian Mean Tracker, BMT; Eqs. 9-11). The β *lesion model* disregards uncertainty and values options solely based on their reward expectations, corresponding to a mean greedy sampling strategy (i.e., $\beta = 0$ in Eq. 6). Finally, the τ *lesion model* replaces the softmax choice function (Eq. 7) with an ϵ-greedy policy (Eq. 12) as an alternative mechanism for random exploration, which samples uniformly from all options with fixed probability ϵ, otherwise selecting the option with the highest UCB value.

The predictive accuracy of models was assessed through leave-one-round-out cross-validation, where predictive performance was quantified by the predicted choice probabilities and log loss across all out-of-sample predictions. For group-level model selection we computed protected exceedance probabilities (pxp), which quantify the probability that a given model is more frequent in the population than all competing models [46]. Across all three groups, the GP-UCB model outperformed all lesioned models, suggesting it as the as the most frequent population model (Fig. 3a).

Additionally, we used a pseudo-R^2 measure (Eq. 13) to quantify predictive accuracy, where $R^2 = 0$ corresponds to the accuracy at chance level and $R^2 = 1$ corresponds to perfect accuracy. Across groups and in line with the pxp analysis, the GP-UCB model outperformed each of the lesioned models (t-tests; all $p < .01$, all $BF \geq 12$; Table S5). In each group, the GP-UCB model was also the most predictive (highest R^2) model for the majority of participants (Fig. 3b).

Overall, these results show that all three components of the GP-UCB model are critical for predicting behavior, aligning with findings from previous studies covering a broad age range in non-clinical populations [29, 35–37, 43].

Levodopa selectively alters uncertainty-directed but not random exploration in PD

The parameters of the GP-UCB model captures distinct facets of learning and exploration.. The parameter λ of the learning model reflects how strongly a learner generalizes, the uncertainty bonus β tracks the level of uncertainty-directed exploration, and the temperature parameter τ indicates the amount of random exploration (Fig. 3c).

Compared to controls, PD− patients had a significantly reduced generalization parameter λ, indicating a deficit in learning and utilizing the spatial regularities of the grid ($U = 832$, $p = .004$, $r_\tau = .28$, $BF = 6.0$). PD+ patients tended to generalize more than PD− patients and less than controls, but the differences were not statistically significant.

The size of the uncertainty bonus markedly differed between groups, with PD− patients showing the highest levels and substantial variability compared to the other groups (Fig. 3b). Both PD+ patients ($U = 271$, $p < .001$, $r_\tau = -.38$, $BF = 19$) and controls ($U = 224$, $p < .001$, $r_\tau = -.44$, $BF > 100$) had lower uncertainty bonuses than patients off medication. PD+ patients and controls did not differ.

We did not observe mean differences in the temperature parameter τ (all $p > .46$) among groups, indicating comparable average levels of random exploration across groups. However, variability was markedly higher in the PD− group than in PD+ patients and controls, similar to the pattern observed for the uncertainty bonus β.

Together, these analyses suggest that the effects of dopaminergic medication are specifically reflected in the degree of uncertainty-directed exploration, with levodopa regulating the "exploration bonus" to levels comparable to those observed in controls without PD. This aligns with findings from a restless bandit paradigm with healthy volunteers, where levodopa reduced the amount of directed exploration, while random exploration remained unaffected by medication.[26]

Controls and PD patients on medication balance directed and undirected exploration better than PD patients off medication

To evaluate how well different parameter settings balance exploration and exploitation, we conducted simulations with the GP-UCB model. In these simulations, we set the amount of generalization to $\lambda = 1$, matching the true smoothness of the reward function in the environments, and systematically varied the amount of random exploration (τ) and the size of the uncertainty bonus (β) to simulate the performance of the GP-UCB model under different levels of directed and random exploration. Figure 3d depicts the resulting performance landscape together with the inferred parameters of all participants. Compared to PD− patients off medication, the parameter estimates for PD+ patients on medication and controls cluster more closely around the optimal

region. Particularly the inflated exploration bonus β places PD− patients in regions yielding low performance.

Discussion

We investigated reward learning and exploratory behavior of PD patients on medication (PD+), PD patients off medication (PD−) and controls in a structured reward-learning task with a large decision space. Because rewards were spatially correlated, learners had to learn and leverage hidden structure to generalize to novel options, which placed specific demands on goal-directed exploration. PD− patients struggled to exploit known high-value options and showed little sensitivity to previous rewards during exploration. A computational model separating random from uncertainty-directed exploration revealed that PD− patients tended towards excessive uncertainty-directed exploration, suggesting impaired regulation of exploratory behavior in a dopamine-depleted state. The performance deficits align with prior evidence of impaired reward learning in PD [7, 8, 11–14, 47–49], although in our task the impairments were more severe. This suggests that structured tasks that require learning about latent regularities are particularly affected in PD, highlighting their potential diagnostic value.

On medication, PD patients largely resembled controls, showing how dopaminergic medication regulates the balance between exploration and exploitation. The restorative effect of levodopa suggests that the observed deficits were specifically caused by dopamine depletion and not other mechanisms underlying PD such as a loss of noradrenergic neurons or neuroinflammatory processes. Notably, while other pharmacological studies in PD [7, 13, 49] also show beneficial effects of levodopa, the improvements in the present study were much stronger.

Behaviorally, the performance deficits of PD− patients result from too much exploration and too little exploitation, consistent with findings in PD patients with apathy [14]. Notably, failures to exploit in our task cannot be attributed to working memory, as previous rewards remained visible. The computational analyses provided additional insight, as goal-directed exploration in the present task reflects two factors: utilizing the spatial correlation in the environment and the consideration of uncertainty in the valuation of choice options. While random exploration was comparable across groups, both generalization and uncertainty-directed exploration were impaired in PD− patients.

First, deficits in generalization were evident in how PD− patients adjusted their search distances based on the reward magnitude of the preceding choice. Leveraging the spatial correlation of rewards should promote local search after high rewards and more distant search after low rewards. Yet, PD− showed only minimal adaption to this structure, which, computationally, resulted in the reduced λ parameter. These results are consistent with a general executive planning deficit in PD [50], and with impairments in model-based learning [10]. Second, we observed a strong increase in the valuation of uncertainty, such that PD− patients placed an overly strong emphasis on uncertainty-directed exploration. The observed increase in uncertainty-guided exploration may seem surprising, as dopamine is often linked to novelty seeking [51]. However, our findings align with recent findings that levodopa in healthy participants attenuated uncertainty-directed exploration in a restless bandit task [26], and that in monkeys a pharmacological inhibition of the dopamine transporter increased their preference for novel options[52]. This combination of lowered reward sensitivity and heightened exploration resembles mechanisms discussed in addiction, where low tonic dopamine is argued to drive drug-seeking to offset reduced reward sensitivity (the *dopamine hypothesis of drug addiction* [53]). Analogously, low tonic dopamine in PD may diminish the salience of known rewards and induce a tendency towards exploring novel or uncertain options. This parallel suggests that a dopamine-depleted state can drive maladaptive shifts in behavioral strategies across distinct clinical conditions.

Our study design differed from previous work in two relevant aspects. First, instead of a healthy control group, we included patients with polyneuropathies. While this control group is not expected to exhibit specific deficits in reward learning or exploration, it limits our ability to conclude that levodopa fully restores performance to the level of healthy individuals. At the same time, an

advantage of our non-healthy control group is that group differences are less likely to be attributable to nonspecific effects of "having a disease". Importantly, all three groups in our study were closely matched on age, gender, depressive symptoms, and also in basic cognitive functioning, which did not differ between PD patients tested on and off medication Second, our withdrawal period in the PD− group was shorter than in most previous studies, which often employ a full overnight withdrawal (e.g., [54, 55]). This limitation arose from the outpatient setting, where minimizing disruption to the patients' regular treatment was a priority. Nevertheless, the robust behavioral effects observed suggest that the shorter withdrawal period did not substantially impact our results.

Conclusion

In sum, our results point to a crucial role of dopamine in the regulation of goal-directed exploration. Conversely, the loss of such regulatory dopaminergic function in PD may account for the marked deficits in reward learning, set shifting [56] and other instrumental activities of daily living that require intact exploratory behavior, such as shopping, transportation, and money management [57–60]. Dopaminergic medication might be an effective therapy in PD to mitigate these impairments, helping patients to navigate the ubiquitous explore-exploit trade-offs they encounter in their daily lives.

Methods

Sample and study design

We tested N=66 adult participants with Parkinson's disease (PD) who regularly receive levodopa (levodopa) for symptomatic treatment. Participants were recruited and tested at a neurologist's outpatient practice; sessions lasted 30–45 minutes. Eligible individuals diagnosed with PD were evaluated based on the Hoehn-Yahr scores recorded in their patient files. The scale assesses disease severity and motor impairments based on a score from 1 to 5, with higher scores indicating greater severity [40]. Recruitment was limited to individuals with scores between 1 and 3, and patients who did not receive antipsychotic medication (except for one patient who received a small dose of 25 mg clozapine to counteract delusional side effects of levodopa).

The study was approved by the Institutional Review Board of the Health and Medical University, Potsdam, Germany. All procedures were carried out in accordance with the Declaration of Helsinki and applicable institutional and national guidelines and regulations. All participants provided written informed consent.

PD patients were randomly assigned to two conditions: testing on medication (PD+) and off medication (PD−). In the PD+ group (N=33), patients' scheduled levodopa dose was administered at least 30 minutes before the start of the experiment. One patient was treated with continuous subcutaneous infusion pump and was therefore assigned to the PD+ group. In the PD− group (N=33), participants were tested immediately before their next scheduled dose, when dopaminergic stimulation from medication was presumably minimal. In sum, the PD+ "on medication" group was tested just after taking levodopa and patients in the PD− "off medication" group were tested just before their next scheduled dose. The behavioral data of one patient in the PD− group was lost due to a computer crash and excluded from the analysisnature maental

The control group (N=35) was recruited in the same practice and consisted of individuals of similar age diagnosed with polyneuropathies, a condition affecting the peripheral nervous system that can lead to physical symptoms such as pain, sensory loss, or motor weakness. However, unlike Parkinson's disease, polyneuropathies typically do not involve central dopaminergic dysfunction or cognitive impairment.

Clinical assessment

We employed standardized measures assessing Parkinson's disease severity, basic cognitive function, and depressive symptoms. PD severity was evaluated using the Hoehn-Yahr scale as documented in patients' most recent clinical records, which rates motor impairments such as postural instability and gait difficulties [40]. Basic cognitive function was assessed through the Mini-Mental State Examination (MMSE) [41]. The test comprises 30 questions pertaining to different domains, including memory (e.g., recalling three objects), temporal and spatial orientation (e.g., date and location), and arithmetic ability. Finally, all participants answered the German version of the Beck Depression Inventory II, a self-report inventory measuring depressive symptoms [39, 61]. No differences on any of these measures was found (Table 1).

Behavioral experiment

Participants performed a reward-based decision-making task based on a spatially-correlated multi-armed bandit task (Fig.1a). Participants completed 10 rounds of the task, each featuring a new environment with the same level of spatial correlation. At the start of each round, one tile was randomly revealed, and participants sequentially sampled 25 tiles. On each trial, they could choose to either click a new tile (explore) or re-click a previously selected tile (exploit). Selections were made by selecting the tile on the computer screen using a mouse, upon which they received a reward in the range [0,50]. Re-clicked tiles showed small variations in reward due to normally distributed noise. Rewards were shown numerically together with a corresponding color, with darker shades of red indicating higher rewards.

The environment in each round was sampled from a pool of 40 distinct environments, which were generated using a radial basis function kernel with length-scale parameter $\lambda = 1$, creating a bivariate reward function on the grid that maps each tile location to a specific reward value. These reward functions gradually varied across the grid, creating environments with spatially-correlated rewards (Fig.1a, right panel). Expressed as a Pearson correlation, the rewards of directly neighboring options correlated at approximately $r \approx 0.6$, decreasing exponentially with spatial distance.

In each of 10 rounds, participants had 25 choices to accumulate rewards. The first round served as a tutorial to familiarize participants with the task, goal, spatial correlation of rewards, possibility of re-clicking tiles and the length of the search horizon. Before the actual task started, each participant had to pass an instruction check with questions pertaining to the instructed goal, that points could be collected both by revealing new tiles and by re-clicking previously revealed tiles, and that points tended to cluster. Data from the tutorial round was excluded from the analyses. The 10th and final round was a bonus round where, after 15 choices, participants were asked to predict rewards for five unrevealed options. Data from this round were also excluded from the main analysis and analyzed separately (SI).

Data analysis

Neither round number (2-9) nor any clinical markers of PD (depressive symptoms, cognitive screening, PD severity) affected performance (Tables S1 - S3). Therefore, these variables were not further considered in the analyses. Data analysis comprised group comparisons, hierarchical regression and correlation analyses. We report frequentist statistics as well as Bayes factors (BFs) to quantify the relative evidence of the alternative hypothesis (H_1) over the null hypothesis (H_0) (see SI for details). To analyze the effect of previous reward on search distance we performed a Bayesian hierarchical regression, including fixed effects for reward, group and the interaction reward $\times$ group, as well as subject-wise random intercepts.

Computational modeling

Gaussian Process Upper Confidence Bound Sampling (GP-UCB) model

The GP-UCB model combines a Bayesian framework for value generalization (GP) with a combination of uncertainty-directed and random exploration (UCB). As such, it provides a comprehensive computational framework for exploratory behavior and how this behavior is shaped by mental health conditions and medication.

Gaussian Process generalization. To model how participants generalize observed rewards across the two-dimensional grid, we use Gaussian Process (GP) regression [62]. A GP is a Bayesian non-parametric model of function learning that has been used as a psychological model of human generalization [34] across spatial [29, 37], abstract [43], social [32, 44], and graph-structured domains [63], predicting both choices and judgments about reward expectations and confidence. Formally, a GP defines a probability distribution over functions mapping inputs to outputs $f : \mathcal{X} \rightarrow Y$. In our case, these functions map grid locations $\mathbf{x} \in \mathcal{X}$ to scalar reward observations $y \in Y$, with the prior distribution taking the form of a multivariate Gaussian:

$$f \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')\big). \tag{1}$$

The GP is fully specified by prior mean function $m(\mathbf{x})$ defining the prior expectations of each input, and a kernel (covariance) $k(\mathbf{x}, \mathbf{x}')$ encoding how strongly rewards at two locations are expected to covary as a function of their distance (see Eq. 2). Without loss of generality, we set the prior mean to zero [62] and use the common radial basis function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right). \tag{2}$$

Here, $\mathbf{x}$ and $\mathbf{x}'$ are the coordinates of two tiles on the grid, and λ is the length-scale parameter, which governs the amount of generalization (i.e., the smoothness of the function). Higher values of λ imply smoother functions, leading to stronger expectations regarding reward correlations. Lower values of λ entail rougher functions, i.e. less correlation among similar options. In our analyses, we treat λ as a free parameter representing the extent to which learners generalize rewards as function of spatial proximity.

To compute posterior predictions for any target location $\mathbf{x}_\star$, we condition the model on a set of observations $\mathcal{D}_t = \{X_t, \mathbf{y}_t\}$ of choices $X_t = [\mathbf{x}_1, \dots \mathbf{x}_t]$ and corresponding reward observations $\mathbf{y}_t = [y_1, \dots, y_t]$ at time t. This posterior also takes the form of a multivariate Gaussian:

$$f(\mathbf{x}_\star) \mid \mathcal{D}_t \sim \mathcal{N}(m(\mathbf{x}_\star|\mathcal{D}_t), v(\mathbf{x}_\star|\mathcal{D}_t)), \tag{3}$$

which is entirely defined by a posterior mean $m(\mathbf{x}_\star|\mathcal{D}_t)$ and a posterior variance $v(\mathbf{x}_\star|\mathcal{D}_t)$. These are computed as:

$$m(\mathbf{x}_\star|\mathcal{D}_t) = \mathbf{k}_\star \big[K_{X,X} + \sigma_\epsilon^2 I\big]^{-1} \mathbf{y}_t \tag{4}$$

$$v(\mathbf{x}_\star|\mathcal{D}_t) = k(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_\star^\top \big[K_{X,X} + \sigma_\epsilon^2 I\big]^{-1} \mathbf{k}_\star. \tag{5}$$

Here, $\mathbf{k}_\star = [k(\mathbf{x}_1, \mathbf{x}_\star), \dots, k(\mathbf{x}_t, \mathbf{x}_\star)]$ is the vector of kernel similarities between past observations and the target location, $K_{X,X}$ is a matrix of pairwise kernel similarities between all past observations in X_t, I is a $t \times t$ identity matrix, and σ_ϵ^2 is the observation noise capturing the stochasticity of reward observations and is fixed to the true reward variability of each arm of the bandit $\sigma_\epsilon^2 = .0001$.

Upper Confidence Bound (UCB) sampling. UCB sampling considers both reward estimates and their uncertainty when valuing options. Formally, this is done by adding an *uncertainty bonus* to the expected rewards of each option:

$$\mathrm{UCB}(\mathbf{x}) = m(\mathbf{x}|\mathcal{D}_t) + \beta\sqrt{v(\mathbf{x}_\star|\mathcal{D}_t)} \tag{6}$$

where the expected reward of an option $m(\mathbf{x}|\mathcal{D}_t)$ captures its exploitation value, and the scaled uncertainty$\beta\sqrt{v(\mathbf{x}_\star|\mathcal{D}_t)}$ captures its exploration value, with β modulating how much exploration is promoted relative to exploitation.

Importantly, balancing rewards and uncertainty requires an appropriate value for β. If β is too small, the learner fails to explore promising but uncertain options, leading to a disproportionate focus on exploitation. Conversely, if β is too large, the uncertainty bonus overrules any reward differences, making all options appear equally attractive and leading to overexploration.

Softmax choice rule. The final step of the model is to translate UCB values into choice probabilities, describing how likely an agent will select each of the 64 options. This is implemented using a softmax function:

$$p(\mathbf{x}) = \frac{\exp(\mathrm{UCB}(\mathbf{x})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{UCB}(\mathbf{x}_j)/\tau)}. \tag{7}$$

The amount of randomness in the choice probabilities is governed by the *temperature parameter* τ. Higher values of τ make the choice probabilities more uniform, such that the choice behavior is less influenced by options' UCB values and more random. Lower value of τ imply that the learner is more sensitive to options' UCB values, making them increasingly likely to be selected. In the limits, if $\tau \to 0$, choice behavior reduces to a greedy mean policy that always selects the option with the highest value (pure exploitation), and if $\tau \to \infty$ all options are chose with equal probability (pure exploration). Here, we treat the temperature parameter τ as a computational marker of a learner's tendency to explore randomly, i.e., in an undirected fashion through inherent decision noise.

Lesioned models

To establish that all components of the GP-UCB model are required to explain behavior, we implemented three lesion variants of the model [37]

The λ lesion model removes the ability to generalize, such that options' rewards are learned independently via a *Bayesian Mean Tracker* (BMT). The BMT is a Kalman filter with time-invariant rewards [64, 65], and as such, can be interpreted as Bayesian variant [66] of the classic Rescorla-Wagner [67] or Q-learning models [68]. Intuitively, reward estimates are updated as a function of prediction error, where the learning rate is dynamically defined based on the degree of uncertainty of the model.

Like the GP, the BMT also assumes a Gaussian prior distribution of reward expectations, but does so independently for each option $\mathbf{x}$:

$$p\big(r_0(\mathbf{x})\big) \sim \mathcal{N}\big(m_0(\mathbf{x}), v_0(\mathbf{x})\big), \tag{8}$$

where $m_0(\mathbf{x}) = 0$ as in the GP, and we set $v_0(\mathbf{x}) = 5$ following [37].

The BMT then computes a posterior distribution of the expected reward for each option, also in the form of a Gaussian, but where the posterior mean $m_t(\mathbf{x})$ and posterior variance $v_t(\mathbf{x})$ are defined independently for each option and computed by the following updates:

$$m_{t+1}(\mathbf{x}) = m_t(\mathbf{x}) + \delta_t(\mathbf{x})G_t(\mathbf{x})\big(y_t(\mathbf{x}) - m_t(\mathbf{x})\big) \tag{9}$$

$$v_{t+1}(\mathbf{x}) = v_t(\mathbf{x})\big(1 - \delta_t(\mathbf{x})G_t(\mathbf{x})\big) \tag{10}$$

Both updates use $\delta_t(\mathbf{x}) = 1$ if option $\mathbf{x}$ was chosen on trial t, and $\delta_t(\mathbf{x}) = 0$ otherwise. Thus, the posterior mean and variance are only updated for the chosen option. The update of the mean is

based on the prediction error $y_t(\mathbf{x}) - m_t(\mathbf{x})$ between observed and anticipated reward, while the magnitude of the update is based on the Kalman gain $G_t(\mathbf{x})$:

$$G_t(\mathbf{x}) = \frac{v_t(\mathbf{x})}{v_t(\mathbf{x}) + \theta_\epsilon^2}, \tag{11}$$

analogous to the learning rate of the Rescorla-Wagner or Q-learning models. Here, the Kalman gain is dynamically defined as a ratio of variance terms, where $v_t(\mathbf{x})$ is the posterior variance estimate and θ_ϵ^2 is the error variance, which we treat as a free parameter and can be interpreted as an inverse sensitivity parameter. Smaller values of θ_ϵ^2 thus result in larger updates of the mean.

The β lesion model evaluates options solely based on their expected rewards, corresponding to a mean-greedy (MG) sampling strategy, and is implemented by setting the uncertainty bonus to $\beta = 0$ (Eq. 6). Effectively, this equates the value of options with their posterior mean $\mathrm{MG}(\mathbf{x}) = m(\mathbf{x}|\mathcal{D}_t)$.

The τ lesion model replaces the softmax choice function (Eq. 7) with an ϵ-greedy policy as an alternative mechanism for random exploration. Under this policy, with probability ϵ a random option is selected and with probability $1 - \epsilon$, the option with the highest UCB value is chosen:

$$p(\mathbf{x}) = \begin{cases} \arg\max \mathrm{UCB}(\mathbf{x}), & \text{with probability } 1 - \epsilon \\ 1/64, & \text{with probability } \epsilon \end{cases} \tag{12}$$

with the parameter ϵ estimated individually for each participant.

Model cross-validation

Models' predictive accuracy was assessed using leave-one-round-out cross-validation based on maximum likelihood estimation [69], with parameter bounds set to the range $[\exp(-5), \exp(4)]$. Specifically, we iteratively held out one round from the task, fitted each model to the remaining seven rounds, and then tested its ability to predict participants' choices on the 25 trials of the holdout round. Predictive accuracy was quantified as the sum of negative log-likelihoods across all out-of-sample predictions. Individual parameter estimates for participants are based on averaging over the cross-validated maximum likelihood estimates.

The negative log-likelihoods served as the model evidence for the hierarchical Bayesian model selection based on protected exceedance probabilities ([46] (Fig.3a), and for quantifying predictive accuracy using a pseudo-R^2 measure, where the summed log loss of each model is compared to a random baseline model. Accordingly, $R^2 = 0$ corresponds to chance performance and $R^2 = 1$ corresponds to theoretically perfect predictions:

$$R^2 = 1 - \frac{\log \mathcal{L}(M_k)}{\log \mathcal{L}(M_{rand})} \tag{13}$$

For participant classification (Fig. 3b) we compared the models within each participant and chose the model with the highest R^2 (or, equivalently, lowest log loss). In a supplementary analysis, we performed a model comparison on the group level likewise using the R^2 measure. Consistent with both hierarchical Bayesian model selection and participant classification, the GP-UCB model achieved the highest R^2 in each group (Table S5 and Fig. S5 in the SI).

Model simulations

To evaluate the performance of the GP-UCB model under different parameter settings we conducted computer simulations (Figure 3d). Specifically, we fixed the amount of generalization to $\lambda = 1$, corresponding to the true smoothness of the reward function in the used environments, and then varied the size of the uncertainty bonus (β) and the amount of random exploration (τ) over logarithmically spaced grids. Both parameters were sampled at 200 values between $\exp(-5)$

(≈ 0.0067) and exp(4) (≈ 54.6), resulting in a total of 40,000 unique (τ, β) combinations. For each parameter combination we simulated 1000 learners searching for rewards using the GP-UCB model, where environments were sampled (with replacement) from the set of 40 environments used in the behavioral study.

In addition, we ran simulations with $\lambda = 0.5$, which is closer to the mean group estimates; these simulations yielded comparable results (Figure S6).

Supplementary information. Supplementary information (SI) accompanies this manuscript and includes additional analyses, figures and tables referenced in the main text.

Declarations

Funding CMW is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.
Conflict of interest/Competing interests The authors declare no competing interests.
Ethics approval and consent to participate The study was approved by the appropriate institutional ethics committee. All participants provided informed consent.
Consent for publication Not applicable.
Data availability All data are available at https://github.com/charleywu/gridsearch_parkinsons.
Materials availability Code for running the behavioral experiment is available at https://github.com/charleywu/gridsearch_parkinsons.
Code availability Analysis code is available at https://github.com/charleywu/gridsearch_parkinsons.
Author contribution BM, MG, CMW and MS conceptualized the study. MS collected the data; a subset contributed to her bachelor's thesis supervised by BM and MG. BM and CMW performed the analyses. BM wrote the first draft of the manuscript. MG, CMW, and MS contributed to writing, review, and editing. All authors read and approved the final manuscript.

Supplementary information

Statistical analyses

Statistical analyses were performed using `R` [70]. We report both frequentist and Bayesian statistics, using Bayes factors (BF) to quantify the relative evidence of the data in favor of the alternative hypothesis (H_1) over the null (H_0). All data and code required for reproducing the statistical analyses and figures are available at https://github.com/charleywu/gridsearch_parkinsons.

For parametric group comparisons, we report paired or independent two-tailed Student's t-tests. For non-parametric comparisons we used the Mann-Whitney U test or Wilcoxon signed-rank test. Bayes factors for the t-tests were computed with the R package `BayesFactor` [71], using its default settings. Bayes factors for rank tests were computed following [72]. Bayesian regressions were performed using the `brms` package [73].

Linear correlations were assessed using Pearson's r, with the Bayes factors computed with the `BayesFactor` package [71], using its default settings. Bayes factors for rank correlations quantified with Kendall's τ were computed using an implementation from [74].

Supplementary behavioral results

No performance differences in relation to round number or clinical indicators

We conducted hierarchical Bayesian regression analyses to examine whether performance differed across task rounds or as a function of clinical indicators.

First, we analyzed participants' mean reward per round as function of round number and group using a hierarchical linear regression. Group, round, and their interaction were specified as population-level ("fixed") effects, with group-level (random) intercepts for participants. This analysis showed only a difference between groups, but no performance differences across rounds (Table S1 and Figure S1).

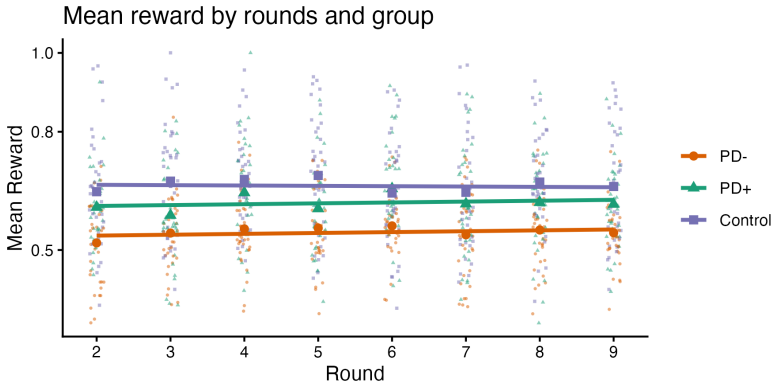

Fig. S1 Reward as a function of rounds. Each line is the population-level effect of a hierarchical Bayesian regression. Large markers show the group means in each round, small dots show individual participants.

Second, we ran a hierarchical Bayesian regression with reward per trial as dependent variable and group, BDI score, and MMSE score as population-level ("fixed") effects, and random intercepts for participants. This analysis also yielded only a substantial population-level effect of group, whereas the influence of both BDI and MMSE was minimal, with posterior estimates close to zero and the 95% HDI including zero (Table S2).

Finally, we ran a hierarchical Bayesian regression for PD patients only, with reward per trial as dependent variable and group, BDI, MMSE, and Hoehn-Yahr score as population-level predictors; and random intercepts for participants. Again, the analysis only yielded an influence of group, i.e.

Table S1 Bayesian linear multilevel regression: mean reward over rounds as a function of group.

	Estimate	**95% HDI**
Population-level effects		
Intercept	0.53	[0.49, 0.57]
Round	0.002	[−0.004, 0.008]
Group PD+	0.08	[0.02, 0.13]
Group Control	0.14	[0.08, 0.18]
Round × Group PD+	0.000004	[-0.008, 0.008]
Round × Group Control	-0.003	[-0.01, 0.01]
Group-level effects		
σ	0.11	[0.11, 0.12]
Observations	808	
N (participants)	101	

Note: Estimates are posterior means with 95% highest density intervals (HDI). Patients on medication (PD+) and controls are compared against PD− patients off medication. σ denotes the residual standard deviation.

Table S2 Bayesian linear multilevel regression: reward as a function of group, BDI, and MMSE.

	Estimate	**95% HDI**
Population-level effects		
Intercept	0.11	[-0.37, 0.57]
Group PD+	0.07	[0.04, 0.10]
Group Control	0.12	[0.09, 0.15]
BDI	0.0002	[-0.003, 0.004]
MMSE	0.02	[-0.001, 0.03]
Group-level effects		
σ	0.25	[0.24, 0.25]
Observations	20000	
N (participants)	100	

Note: Estimates are posterior means with 95% highest density intervals (HDI). Patients on medication (PD+) and controls are compared against PD− patients off medication. σ denotes the residual standard deviation. BDI and MMSE were entered as continuous predictors. One participant in the control group had a missing MMSE score and was therefore excluded from the analysis.

being tested on or off medication (Table S2). The coefficients for all clinical measures were very small, with all posterior means < 0.01 and their HDIs including zero.

In sum, in none of the analyses round number or clinical variables were related to performance. We therefore averaged rewards across rounds and did not further include the clinical indicators in the analyses.

Bonus round: Reward predictions, confidence, and choices

The 10th and last round of the task as a "bonus round" in which participants first made 15 search decisions as in the previous rounds and then predicted the rewards for 5 randomly chosen, previously unobserved tiles. For each tile, they also indicated how confident they were in their prediction (0-10). Subsequently, they chose one of the five tiles and continued the round as usual until the search horizon was exhausted.

Prediction errors were measured as the absolute difference between a participant's predictions and the mean of the true underlying Gaussian distribution. For each participant, we computed the

Table S3 Bayesian linear multilevel regression including only patients with PD: reward as a function of group, BDI, MMSE, and Hoehn–Yahr score.

	Estimate	**95% HDI**
Population-level effects		
Intercept	0.31	[-0.22, 0.85]
Group PD+	0.07	[0.05, 0.10]
BDI	0.002	[-0.001, 0.006]
MMSE	0.007	[-0.01, 0.03]
Hoehn–Yahr	0.006	[-0.01, 0.03]
Group-level effects		
σ	0.24	[0.24, 0.25]
Observations	12200	
N (participants)	66	

Note: Estimates are posterior means with 95% highest density intervals (HDI). Patients on medication (PD+) are compared against PD− patients off medication. σ denotes the residual standard deviation. BDI, MMSE, and Hoehn–Yahr scores were entered as continuous predictors.

mean absolute error by averaging over the five predictions. All three groups had lower prediction error than a random baseline (Figure S2; all $p < .001$). Controls had lower prediction error than PD− patients off medication ($t(67) = -2.7$, $p = .009$, $d = 0.6$, $BF = 5.0$). PD+ patients where slightly worse than controls and slightly better than PD− patients, but did not differ from either group (both $p > .19$).

Across all participants and judgments, there was no relation between participants' prediction error and confidence in their judgments ($r = -.05$). A Bayesian hierarchical regression ()Table S4 with confidence and group as population-level ("fixed") effects, and subject-wise random intercept showed only weak and inconsistent relationships across groups (Figure S3). The control group showed a weak negative relation (i.e., higher confidence was associated with lower prediction error), the PD+ group a weak positive trend (i.e., lower confidence was associated with higher prediction error), and the PD− group showed no association.

Table S4 Bayesian multilevel regression: prediction error as a function of group, confidence, and their interaction.

	Estimate	**95% HDI**
Population-level effects		
Intercept	0.180	[0.151, 0.209]
Confidence	-0.002	[-0.008, 0.004]
Group PD+	-0.039	[-0.085, 0.008]
Group Control	-0.001	[-0.052, 0.048]
Confidence × Group PD+	0.005	[-0.004, 0.014]
Confidence × Group Control	-0.007	[-0.018, 0.002]
Group-level effects		
σ	0.115	[0.107, 0.122]
Observations	515	
N (participants)	103	

Note: Estimates are posterior means with 95% highest density intervals (HDI). σ denotes the residual standard deviation. The model includes interactions between group and confidence (howSecure).

After making reward predictions for the five tiles, participants could choose one of them and then continued the round as usual. (Figure S4) compares the predicted rewards for chosen and

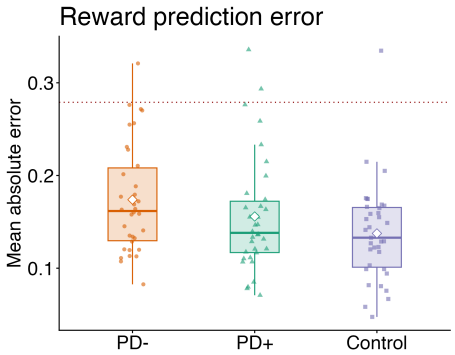

Fig. S2 Prediction error between participants' estimates and the true underlying expected rewards in the bonus round. For reference, the dashed line represents the expected error for a randomly choosing learner.

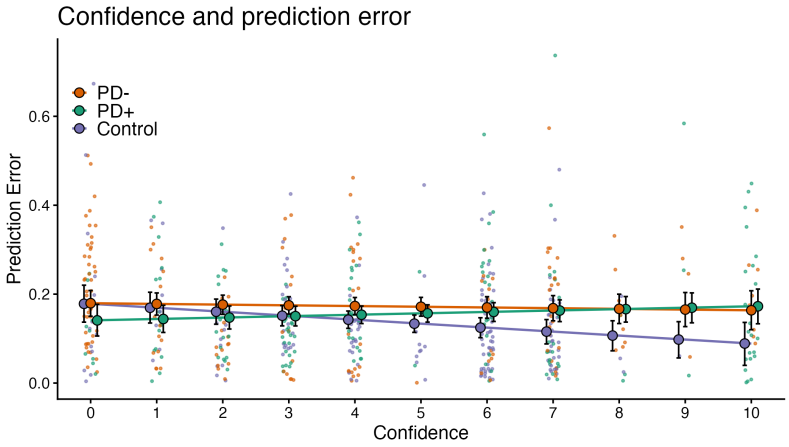

Fig. S3 Predictions of a Bayesian hierarchical regression with prediction error as a function of confidence and group. Large dots show posterior means, with error bars denoting 95% credible intervals. Small dots show observed individual data points.

unchosen tiles, where the predicted reward for unchosen tiles is based on averaging across all four tiles that were not chosen. In the control group, predicted rewards for chosen tiles were higher than for unchosen tiles ($t(34) = 2.9$, $p = .007$, $d = 0.6$, $BF = 6.1$), as well as in PD+ patients ($t(33) = 2.1$, $p = .047$, $d = 0.3$, $BF = 1.2$). No difference was obtained for PD− patients off medication ($p > .8$). In none of the groups average confidence differed between chosen and unchosen tiles (all $p > .3$).

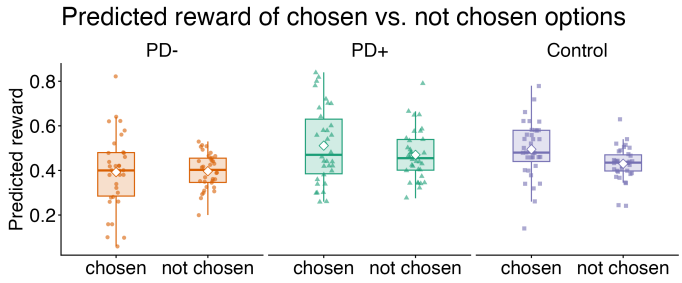

Fig. S4 Participants' reward predictions for chosen versus not chosen options in the bonus round. Small dots show individual data points.

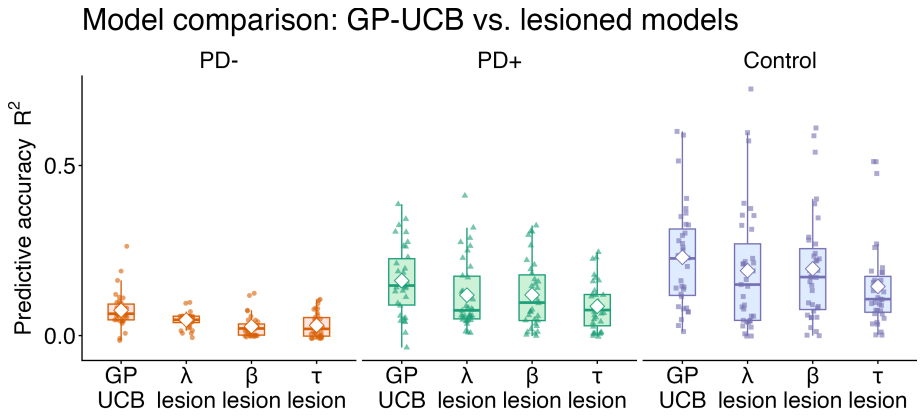

Fig. S5 Predictive model accuracy. Small dots show individual data points.

Supplementary computational modeling results

Model comparison: All components of the GP-UCB model are critical to explain behavior

The predictive accuracy of the computational models was evaluated through leave-one-round-out cross-validation using maximum likelihood estimation [69]. Predictive accuracy was quantified as the sum of negative log-likelihoods across all out-of-sample predictions.

In addition to the hierarchical Bayesian model selection (Fig. 3a) and individual participant classification (Fig. 3b), we also compared the predictive accuracy R^2 across models and groups. In each group, the GP-UCB model achieved the highest predictive accuracy compared to the three lesioned models (Fig. S5). Table S5 shows the results of t-tests comparing the mean R^2 of the GP-UCB model to each lesioned model, showing that the GP-UCB model consistently outperformed all lesioned variants across groups.

Simulated performance of the GP-UCB model

To evaluate how well different parameter settings balance exploration and exploitation, we conducted simulations with the GP-UCB model. Figure 3d in the main text shows participant parameters in relation to the performance landscape with the length-scale parameter of the RBF kernel (Eq. 2) set to $\lambda = 1$, the true amount of spatial correlation in the used environments.

We additionally conducted simulations with $\lambda = 0.5$, which is closer to the mean group estimates of λ, $M_{\text{PD}-} = 0.53$, $M_{\text{PD}+} = 0.56$, $M_{\text{Control}} = 0.64$. Again, we systematically varied the amount of

Group	Comparison	n	t	p	d	BF
PD−	λ lesion	34	3.8	$p < .001$	0.7	$BF = 47$
PD−	β lesion	34	5.8	$p < .001$	1.1	$BF > 100$
PD−	τ lesion	34	5.2	$p < .001$	1.0	$BF > 100$
PD+	λ lesion	34	3.4	$p = .002$	0.4	$BF = 18$
PD+	β lesion	34	3.8	$p < .001$	0.4	$BF = 58$
PD+	τ lesion	34	8.7	$p < .001$	0.9	$BF > 100$
Control	λ lesion	35	3.2	$p = .003$	0.2	$BF = 12$
Control	β lesion	35	3.5	$p = .001$	0.2	$BF = 28$
Control	τ lesion	35	7.9	$p < .001$	0.6	$BF > 100$

Table S5 Pairwise comparisons of R^2 values between GP-UCB model and lesioned variants across groups.

random exploration (τ) and the size of the uncertainty bonus (β) over logarithmically spaced grids. Both parameters were sampled at 200 values between $\exp(-5)$ (≈ 0.0067) and $\exp(4)$ (≈ 54.6), resulting in a total of 40,000 unique (τ, β) combinations. For each parameter combination we simulated 1000 learners searching for rewards using the GP-UCB model, where environments were sampled (with replacement) from the set of 40 environments used in the behavioral study.

Figure S6 depicts the resulting performance landscape together with participants' parameter estimates for β and τ. The results are similar to performance landscape for $\lambda = 1$: Parameter values of controls and PD+ patients are closer to the optimal region compared to those of PD− patients off medication. For the latter, especially the too high β valued shift parameter estimates into low-reward regions of the landscape.

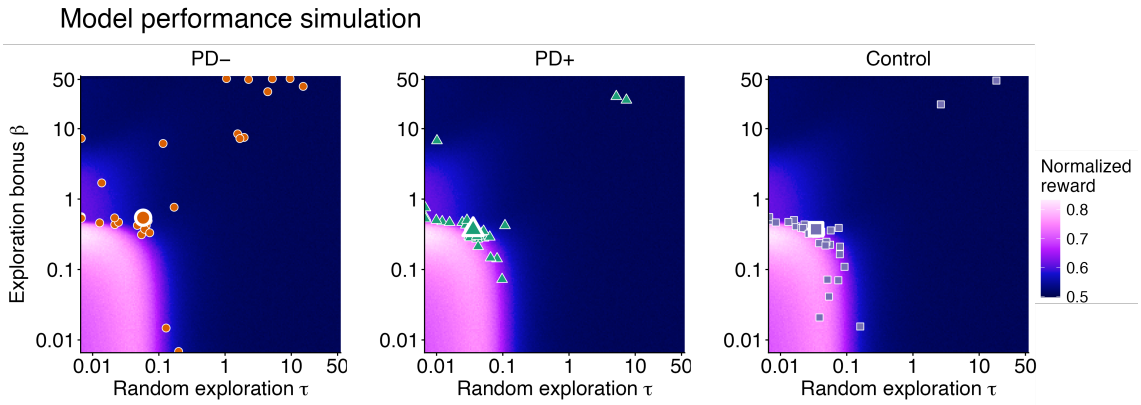

Fig. S6 Performance landscape of the GP-UCB model across different amounts of random exploration τ and uncertainty-directed exploration β. The amount of generalization was fixed to λ=0.5 and β and τ were varied, with 1000 simulated learners per parameter combination. Each dot shows one participant; the larger markers indicate the group median.

References

[1] Hornykiewicz, O. The discovery of dopamine deficiency in the parkinsonian brain. *Journal of Neural Transmission. Supplementum* 9–15 (2006).

[2] Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).

[3] Schultz, W. Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* **57**, 87–115 (2006).

[4] Glimcher, P. W. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences* **108**, 15647–15654 (2011).

[5] Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience* **12**, 1062–1068 (2009).

[6] Kayser, A. S., Mitchell, J. M., Weinstein, D. & Frank, M. J. Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology* **40**, 454–462 (2015).

[7] Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. Cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).

[8] Peterson, D. A. *et al.* Probabilistic reversal learning is impaired in parkinson's disease. *Neuroscience* **163**, 1092–1101 (2009).

[9] Cools, R., Lewis, S. J., Clark, L., Barker, R. A. & Robbins, T. W. L-dopa disrupts activity in the nucleus accumbens during reversal learning in parkinson's disease. *Neuropsychopharmacology* **32**, 180–189 (2007).

[10] Sharp, M. E., Foerde, K., Daw, N. D. & Shohamy, D. Dopamine selectively remediates 'model-based' reward learning: a computational approach. *Brain* **139**, 355–364 (2016).

[11] Cools, R., Altamirano, L. & D'Esposito, M. Reversal learning in parkinson's disease depends on medication status and outcome valence. *Neuropsychologia* **44**, 1663–1673 (2006).

[12] Frank, M. J., Samanta, J., Moustafa, A. A. & Sherman, S. J. Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science* **318**, 1309–1312 (2007).

[13] Palminteri, S. *et al.* Pharmacological modulation of subliminal learning in parkinson's and tourette's syndromes. *Proceedings of the National Academy of Sciences* **106**, 19179–19184 (2009).

[14] Gilmour, W. *et al.* Impaired value-based decision-making in parkinson's disease apathy. *Brain* **147**, 1362–1376 (2024).

[15] Seymour, B. *et al.* Deep brain stimulation of the subthalamic nucleus modulates sensitivity to decision outcome value in parkinson's disease. *Scientific Reports* **6**, 32509 (2016).

[16] Djamshidian, A., O'Sullivan, S. S., Wittmann, B. C., Lees, A. J. & Averbeck, B. B. Novelty seeking behaviour in parkinson's disease. *Neuropsychologia* **49**, 2483–2488 (2011).

[17] Mehlhorn, K. *et al.* Unpacking the exploration–exploitation tradeoff: a synthesis of human and animal literatures. *Decision* **2**, 191 (2015).

[18] Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D. & Couzin, I. D. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences* **19**, 46–54 (2015).

[19] Cohen, J. D., McClure, S. M. & Yu, A. J. 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* **362**, 933–942 (2007).

[20] Gittins, J. C. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **41**, 148–164 (1979).

[21] Reverdy, P. B., Srivastava, V. & Leonard, N. E. Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE* **102**, 544–571 (2014).

[22] Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology* **53**, 168–179 (2009).

[23] Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**, 397–422 (2002).

[24] Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science* **7**, 351–367 (2015).

[25] Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* **143**, 2074—2081 (2014).

[26] Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F. & Peters, J. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife* **9**, e51260 (2020).

[27] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).

[28] Hertwig, R., Barron, G., Weber, E. U. & Erev, I. Decisions from experience and the effect of rare events in risky choice. *Psychological Science* **15**, 534–539 (2004).

[29] Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour* **2**, 915–924 (2018).

[30] Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**, 927–943 (2018).

[31] Schulz, E. *et al.* Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences* **116**, 13903–13908 (2019).

[32] Wu, C. M. *et al.* Adaptive mechanisms of social and asocial learning in immersive foraging environments. *Nature Communications* **16**, 3539 (2025).

[33] Gershman, S. J., Malmaud, J. & Tenenbaum, J. B. Structured representations of utility in combinatorial domains. *Decision* **4**, 67–86 (2017).

[34] Wu, C. M., Meder, B. & Schulz, E. Unifying principles of generalization: Past, present, and future. *Annual Review of Psychology* **76**, 275–302 (2025).

[35] Meder, B., Wu, C. M., Schulz, E. & Ruggeri, A. Development of directed and random exploration in children. *Developmental Science* **24**, e13095 (2021).

[36] Schulz, E., Wu, C. M., Ruggeri, A. & Meder, B. Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science* **30**, 1561–1572 (2019).

[37] Giron, A. P. *et al.* Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour* **7**, 1955–1967 (2023).

[38] Vellani, V., de Vries, L. P., Gaule, A. & Sharot, T. A selective effect of dopamine on information-seeking. *eLife* **9**, e59152 (2020).

[39] Beck, A. T., Steer, R. A. & Brown, G. Beck depression inventory–II (BDI-II) [database record]. APA PsycTests (1996). American Psychological Association.

[40] Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression, and mortality. *Neurology* **17**, 427–427 (1967).

[41] Folstein, M. F., Folstein, S. E. & McHugh, P. R. "mini-mental state"': a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* **12**, 189–198 (1975).

[42] Jost, S. T. *et al.* Levodopa Dose Equivalency in Parkinson's Disease: Updated Systematic Review and Proposals. *Movement Disorders: Official Journal of the Movement Disorder Society* **38**, 1236–1252 (2023).

[43] Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Computational Biology* **16**, e1008149 (2020).

[44] Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W. & Wu, C. M. Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences* **121**, e2404928121 (2024).

[45] Shohamy, D., Myers, C. E., Geghman, K. D., Sage, J. & Gluck, M. A. L-dopa impairs learning, but spares generalization, in parkinson's disease. *Neuropsychologia* **44**, 774–784 (2006).

[46] Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage* **84**, 971–985 (2014).

[47] Bódi, N. *et al.* Reward-learning and the novelty-seeking personality: a between- and within-subjects study of the effects of dopamine agonists on young parkinson's patients. *Brain* **132**, 2385–2395 (2009).

[48] Rutledge, R. B. *et al.* Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience* **29**, 15104–15114 (2009).

[49] van Nuland, A. J. *et al.* Effects of dopamine on reinforcement learning in parkinson's disease depend on motor phenotype. *Brain* **143**, 3422—3434 (2020).

[50] Dirnberger, G. & Jahanshahi, M. Executive dysfunction in parkinson's disease: A review. *Journal of Neuropsychology* **7**, 193–224 (2013).

[51] Dellu, F., Piazza, P. V., Mayo, W., Le Moal, M. & Simon, H. Novelty-seeking in rats-biobehavioral characteristics and possible relationship with the sensation-seeking trait in man. *Neuropsychobiology* **34**, 136–145 (1996).

[52] Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Dopamine modulates novelty seeking behavior during decision making. *Behavioral Neuroscience* **128**, 556–566 (2014).

[53] Melis, M., Spiga, S. & Diana, M. The Dopamine Hypothesis of Drug Addiction: Hypodopaminergic State **63**, 101–154 (2005).

[54] Cools, R., Barker, R. A., Sahakian, B. J. & Robbins, T. W. Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cerebral Cortex* **11**, 1136–1143 (2001).

[55] McCoy, B., Jahfari, S., Engels, G., Knapen, T. & Theeuwes, J. Dopaminergic medication reduces striatal sensitivity to negative outcomes in parkinson's disease. *Brain* **142**, 3605–3620 (2019).

[56] Cools, R., Barker, R. A., Sahakian, B. J. & Robbins, T. W. Mechanisms of cognitive set flexibility in parkinson's disease. *Brain* **124**, 2503–2512 (2001).

[57] Shulman, L. M. *et al.* The evolution of disability in parkinson disease. *Movement Disorders* **23**, 790–796 (2008).

[58] Young, T. L., Granic, A., Yu Chen, T., Haley, C. B. & Edwards, J. D. Everyday reasoning abilities in persons with parkinson's disease. *Movement Disorders* **25**, 2756–2761 (2010).

[59] Martin, R. C. *et al.* Impaired financial abilities in parkinson's disease patients with mild cognitive impairment and dementia. *Parkinsonism & Related Disorders* **19**, 986–990 (2013).

[60] Foster, E. R. Instrumental activities of daily living performance among people with parkinson's disease without dementia. *The American Journal of Occupational Therapy* **68**, 353–362 (2014).

[61] Hautzinger, M., Keller, F. & Kühner, C. *Beck depressions-inventar (BDI-II)* (Harcourt Test Services, 2006).

[62] Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* (MIT Press Cambridge, MA, 2006).

[63] Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. *Computational Brain & Behavior* **4**, 125–147 (2021).

[64] Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nature Neuroscience* **3**, 1218–1223 (2000).

[65] Wu, C. M., Schulz, E., Pleskac, T. J. & Speekenbrink, M. Time pressure changes how people explore and respond to uncertainty. *Scientific Reports* **12**, 1–14 (2022).

[66] Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput Biol* **11**, e1004567 (2015).

[67] Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* **2**, 64–99 (1972).

[68] Watkins, C. J. & Dayan, P. Q-learning. *Machine Learning* **8**, 279–292 (1992).

[69] Mullen, K. M., Ardia, D., Gil, D. L., Windover, D. & Cline, J. Deoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software* **40**, 1–26 (2011).

[70] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2025). URL https://www.R-project.org/.

[71] Morey, R. D. & Rouder, J. N. *BayesFactor: Computation of Bayes Factors for Common Designs* (2024). URL https://CRAN.R-project.org/package=BayesFactor. R package version 0.9.12-4.7.

[72] van Doorn, J., Ly, A., Marsman, M. & Wagenmakers, E.-J. Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and spearman's ρ. *Journal of Applied Statistics* **47**, 2984–3006 (2020).

[73] Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).

[74] van Doorn, J., Ly, A., Marsman, M. & Wagenmakers, E.-J. Bayesian inference for kendall's rank correlation coefficient. *The American Statistician* **72**, 303–308 (2018).