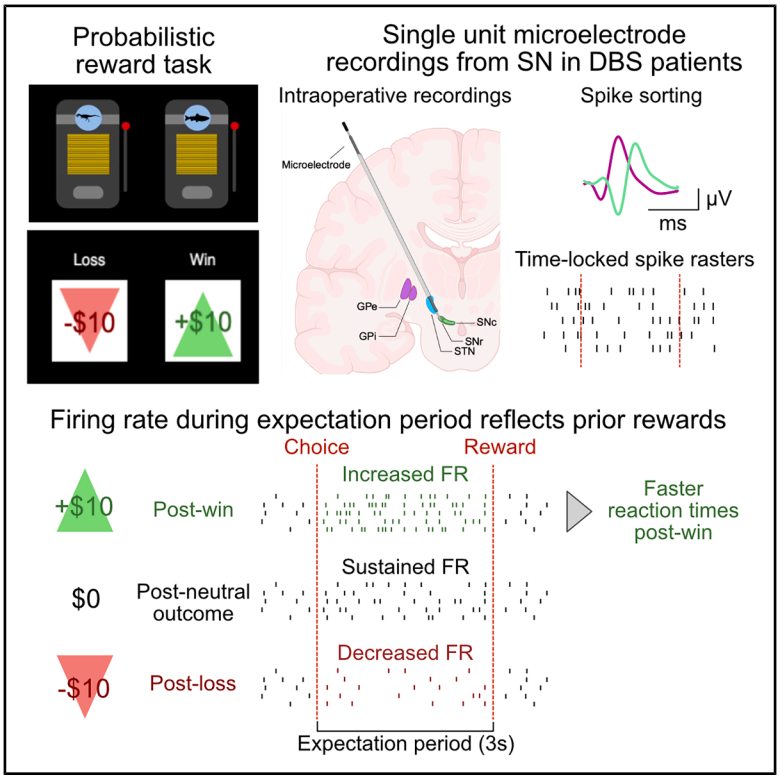
iScience

Article

# Sustained activity of human substantia nigra neurons reflect prior rewards

## Graphical abstract




## Highlights

- Sustained firing of human substantia nigra neurons increases after prior reward
- Prior reward history is also associated with faster subsequent reaction times
- The results link neuronal responses to recent reward history and motivational vigor

## Authors

Zarghona Imtiaz, Ayaka Kato, Brian H. Kopell, ..., Alexander W. Charney, Xiaosi Gu, Ignacio Saez

## Correspondence

ignacio.saez@mssm.edu

## In brief

Health sciences

Imtiaz et al., 2026, iScience *29*, 115458
April 17, 2026 
https://doi.org/10.1016/j.isci.2026.115458

CellPress

iScience

CellPress
OPEN ACCESS

Article

# Sustained activity of human substantia nigra neurons reflect prior rewards

Zarghona Imtiaz,[1] Ayaka Kato,[2,6] Brian H. Kopell,[3,4,6] Salman E. Qasim,[2] Arianna Neal Davis,[1] Lizbeth Nunez Martinez,[1] Matthew Heflin,[2] Kaustubh Kulkarni,[2] Amr Morsi,[3,4] Alexander W. Charney,[2,5] Xiaosi Gu,[1,2,3] and Ignacio Saez[1,3,4,7,*]

[1]Nash Family Department of Neuroscience and the Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[3]Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5]Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: ignacio.saez@mssm.edu
https://doi.org/10.1016/j.isci.2026.115458

## SUMMARY

Dopamine (DA) signals from substantia nigra (SN) neurons encode reward prediction errors (RPEs) and have been implicated in motor control, reward processing, and motivational vigor. However, how recent reward history, related to reward expectations, is represented within the dopaminergic system remains poorly understood, particularly in humans, due to the difficulty of recording DA neuron activity directly. To address this, we performed single-unit recordings from the SN of patients with Parkinson's disease undergoing neurosurgery while they played a two-armed bandit decision-making task. We found that the firing rates (FRs) of putative DA neurons during reward expectation were modulated by previous trial outcomes, with higher FRs following positive outcomes. This increase in FRs was associated with faster subsequent reaction times (RTs), suggesting a link between neural signals reflecting prior reward and behavioral response vigor. These results provide a potential physiological substrate for how reward history influences behavior through the modulation of human dopaminergic activity.

## INTRODUCTION

The dopaminergic system is crucial for reward-based learning, motivational vigor, and other aspects of goal-directed behavior. The activity of individual dopamine (DA) neurons in the ventral tegmental area (VTA) and substantia nigra (SN) encodes reward prediction errors (RPEs).[1,2] RPEs reflect the difference between actual and expected reward and constitute an efficient and robust signal underlying behavioral adaptation in uncertain environments. Abundant evidence across species[1,3–12] supports the notion that dopaminergic activity, including the spiking rates of individual DA neurons, reflects RPEs. However, how reward expectations themselves—the essential baseline against which RPEs are computed—are represented within the dopaminergic system remains less well understood.

Studies in animal models have found electrophysiological correlates of reward expectation in various brain regions, including the caudate nucleus,[13] striatum,[9] orbitofrontal cortex (OFC),[14] and globus pallidus externus (GPe)[15]; DA neurons in the rodent VTA then combine information about reward outcomes and expectations into a largely homogeneous RPE signal which is broadcast to downstream targets.[12,16] Conversely, DA neuron activity during the expectation period correlates with the expected value of the upcoming action.[17,18] Competing models of dopaminergic signaling make different predictions regarding the locus of this integration. If reward outcomes and expectations are combined upstream, DA neurons would primarily reflect RPEs; if integration occurs within DA neurons themselves, both expectation- and outcome-related signals could be separable in DA activity.

This knowledge gap of how DA neurons integrate information about prior rewards and reward expectation and reward into RPEs is particularly salient in the human literature, due to difficulties of recording activity from DA regions in living humans. Previous studies have leveraged neurosurgical interventions to examine human dopaminergic activity using microelectrode recordings to record single unit activity from the SN,[6,7] or voltammetry approaches to record DA transients in the striatum.[19–21] These studies have demonstrated that post-reward activity in the human dopaminergic system reflects reward signals, including unexpected rewards and RPEs. However, whether and how reward expectation signals, reflecting reward history, short-term expectations, or incremental learning mechanisms (RPEs) are encoded in the human dopaminergic system remains unknown. In this study, we use the term "expectation signal" to refer specifically to neural activity during the pre-outcome period that reflects recent reward history (i.e., the outcome of the immediately preceding trial), rather than a formal predicted value derived from incremental reinforcement learning models.

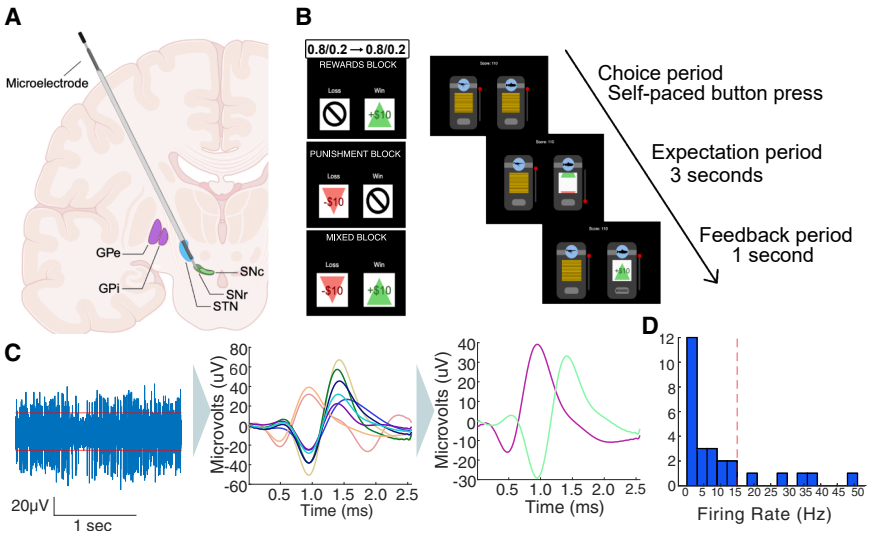



**Figure 1. Invasive microelectrode recordings and the slot machine task**

(A) Electrode placement during DBS surgery for patients with Parkinson's disease. Microelectrodes were parked past the ventral border of the subthalamic nucleus, and single-unit recordings were taken from the substantia nigra.

(B) Two-armed bandit slot machine behavioral task. Patients choose between two slot machines with 80%/20% and 20%/80% probability of a better/worse outcome. Patients play 3×35-trial blocks in pseudo-randomized order, with either +$10/0 (reward block), +$10/-$10 (mixed block), or -$10/$0 (punishment block) outcomes. The contingencies switched twice between slot machines during each block. The patients are shown the slot machines until they make a choice. After a machine is chosen, it spins for 3 s until landing on the outcome, which stays on the screen for 1 s before the slot machines reset and a new trial starts.

(C) Microelectrode data (left) is high-pass filtered, spike sorted to identify individual units (middle), and manually processed to discard or merge clusters into the final units per recording (right).

(D) Histogram shows the firing rates of all neurons collected. The red dashed line represents 15 Hz, which was used as the firing rate cutoff between putative dopaminergic (<15 Hz) and GABAergic (≥15 Hz) neurons. We also used a waveform width cutoff (>0.8 ms for putative dopaminergic neurons) but found no neurons that were classified as putative GABAergic neurons under this criterion (all >0.8 ms).

Here, we sought to investigate whether reward expectation signals can be detected in the activity of individual human SN neurons. We recorded single-unit activity (SUA) intraoperatively from the SN of patients undergoing deep-brain stimulation (DBS) surgery for Parkinson's disease (PD) while the patient played a two-armed bandit task in which they chose between two slot machines with different outcome probabilities that reversed at certain points in the task. We examined neuronal firing rates (FRs) in combination with the computational modeling of patients' strategies. Our results indicate that patients' choices were adequately modeled by lose-switch/win-stay or reward learning (Rescorla-Wagner (RW)) models, in which choices are strongly determined by the outcome of the preceding trial. Previous trial outcome was reflected in neural activity: The FR of putative DA SN neurons was higher following positive (+$10) compared to neutral ($0) and negative (-$10) outcomes. Furthermore, previous outcomes impacted patients' reaction times (RTs), with faster responses following positive compared to negative/neutral outcomes. Together, these results suggest that the neural activity of human SN neurons reflects recent reward history and could be associated with variation in behavioral response vigor. Our results provide novel evidence for the potential encoding of reward history within the human dopaminergic system and suggest that this activity may mediate the relationship between reward history and motivational drive.

## RESULTS

We carried out SUA recordings during two-armed bandit play in 13 sessions from patients with PD ($n$ = 11; 7 male, 4 female, average age = 65.71 ± 7.69 years) who were undergoing DBS lead implantation in the subthalamic nucleus (STN). The microelectrode tip was placed in the SN, past the ventral border of the STN (see Figure 1A and Methods) to record activity of SN neurons. We carried out 8 recordings from the left side of the brain and 5 recordings from the right side of the brain, with bilateral recordings in two patients (Table S1). The location of the electrodes was calculated in patient space as opposed to MNI space to account for errors associated with co-registering small subcortical structures. We recorded from a total of 27 neurons, for an average of 2.077 ± 0.277 neurons per session. To determine the putative identity of individual neurons, we used two different criteria: FR and spike width. We classified neurons with an FR below 15 Hz and a waveform width >0.8 ms as putative DA neurons and those with greater than 15 Hz FR and/or a waveform width <0.8 ms as putative GABAergic interneurons.[22] Out of our 27 neurons, 22 were classified as putative DA neurons (average FR = 4.883 ± 4.657 Hz) and 5 neurons were categorized as GABAergic (average FR = 33.414 ± 10.607 Hz, Figure 1D, example peristimulus time histograms for example neurons in Figure S1).

Patients played a multi-round two-armed bandit task in which they chose between two slot machines with probabilistic reward outcomes (Figure 1B) with the goal of maximizing their overall reward. The task had three blocks with 35 trials per block: a reward block (+$10/$0 outcomes), a punishment block (-$10/$0 outcomes), and a mixed block (+$10/-$10 outcomes). Block order was randomized across patients. One of the machines had a high probability (80%) of resulting in the better outcome, and the other had a low probability (20%) (Figure 1B). To minimize motor confounds, the initial location (left/right) of the high reward machine was randomized, and reward contingencies were reversed twice throughout the task, between trials 12/13 and trials 24/25.

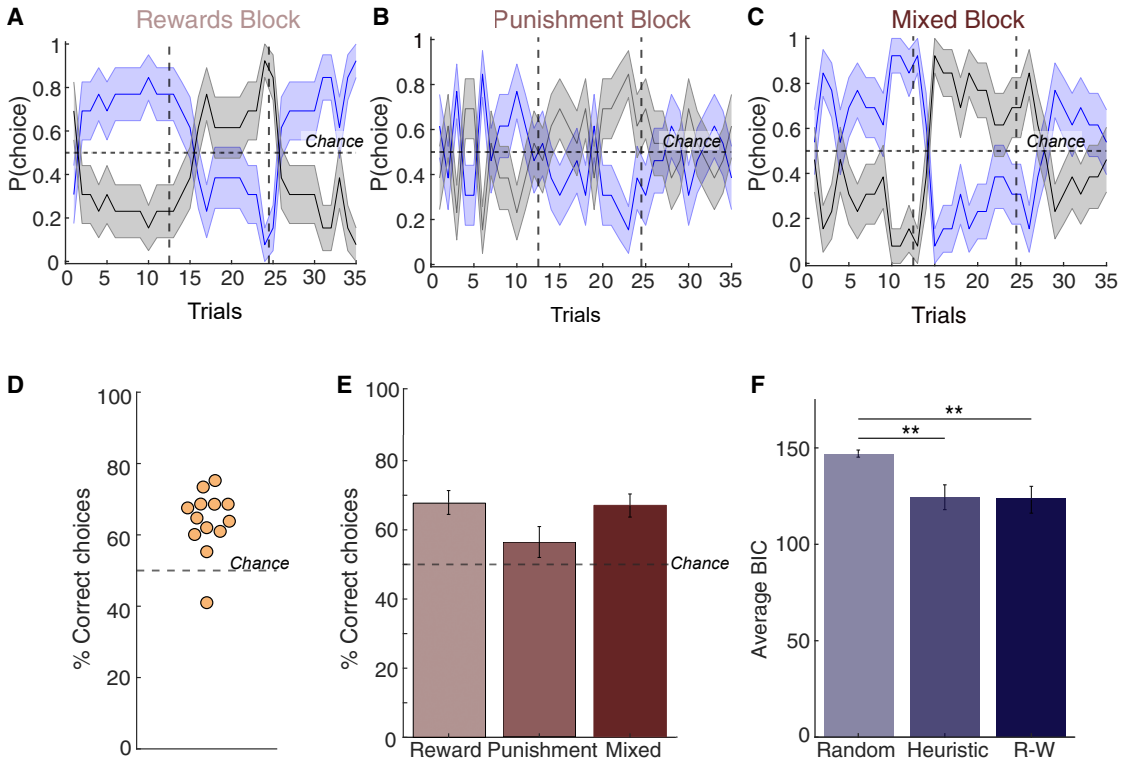



**Figure 2. Patient behavior and computational modeling**

Average patient behavior across the (A) rewards, (B) punishment, and (C) mixed blocks. Patients played 3 blocks corresponding to reward/punishment/mixed conditions in pseudo-randomized order. Within each block, win/loss probabilities (80%/20%) were switched twice (dashed vertical lines). The horizontal dashed line represents chance (50%).

(D and E) Percent of correct choices in which the patient picked the machine with more frequent better outcomes, per patient, across all blocks (D) and the average ± standard error of the mean within each block (E). Dashed line represents the chance (50%). The majority of patients selected the correct machine over 50% of the time ($n$ = 10/11 patients).

(F) Computational modeling of behavior compares the three models considered: a random choice model (Random), a Heuristic win-stay/lose-switch model (Heuristic), and a Rescorla-Wagner model (Rescorla Wagner, RW). The RW produced the best fit to patient behavior (lowest average BIC score), but there was no significant difference between the average BIC scores of the Heuristic and RW models (paired $t$ test, $p$-value = 0.7225). Both Heuristic and RW models had a significantly lower BIC score than the random model (paired $t$ test; heuristic $p$-value = 0.0046, Rescorla-Wagner $p$-value = 0.0049). Average BIC ± standard erorr of the mean shown across models.

## Computational model reveals patients' choices were dependent on the outcome of previous trials

We started by characterizing patient behavior using model-free metrics: number of wins and losses per block, as well as the percentage of trials in which the patient picked the machine with the higher reward probability (correct choices). Patients adjusted their behavior after contingency reversals and picked the new machine with the higher reward rate (Figures 2A–2C, reward = 67.91 ± 12.29%, mixed = 67.03 ± 12.0%, punishment = 56.57 ± 16.0%). Overall, patients performed better than chance (Figure 2D, $p$ = 3.6583e-11, binomial test), indicating that they understood the game. There was no difference in the proportion of correct choices between blocks (Figure 2E, reward vs. punishment: $p$ = 0.081, t(12) = 1.905, CI = [-24.502 1.645]; reward vs. mixed: $p$ = 0.815, t(12) = 0.240, CI = [-8.866 7.108]; punishment vs. mixed: $p$ = 0.060, t(12) = −2.078, CI = [-21.613 0.515], paired $t$ test).

Next, we sought to further understand the behavioral strategies used by patients through the computational modeling of behavior. We fit patient behavior to three different model alternatives: a random model in which patients choose randomly between the two options, a heuristic win-stay/lose-shift model, and a RW model commonly used in RL.[23–25] In the heuristic model, actions are determined by the previous trial outcome (either a "good" or a "bad" outcome, which varied depending on block type), using a win-stay/lose-shift rule. Under the RW model, the agent learns from the previous outcomes to determine the values of each slot machine and act accordingly. To determine which model fit the data the best, we focused on the BIC score of the model, which balances model fit with a penalty for the number of parameters and therefore helps to avoid overfitting and facilitates fair comparison between models with different complexities. The RW model produced the lowest average BIC score when fitting the patients' behavior overall, closely followed by the heuristic win-stay/lose-shift model (Figure 2F; Figure S2 for the result for parameter recovery). The estimated learning rate for the RW model for most patients ($n$ = 7/13) was 1 (mean = 0.7390, std = 0.4110), indicating that patient

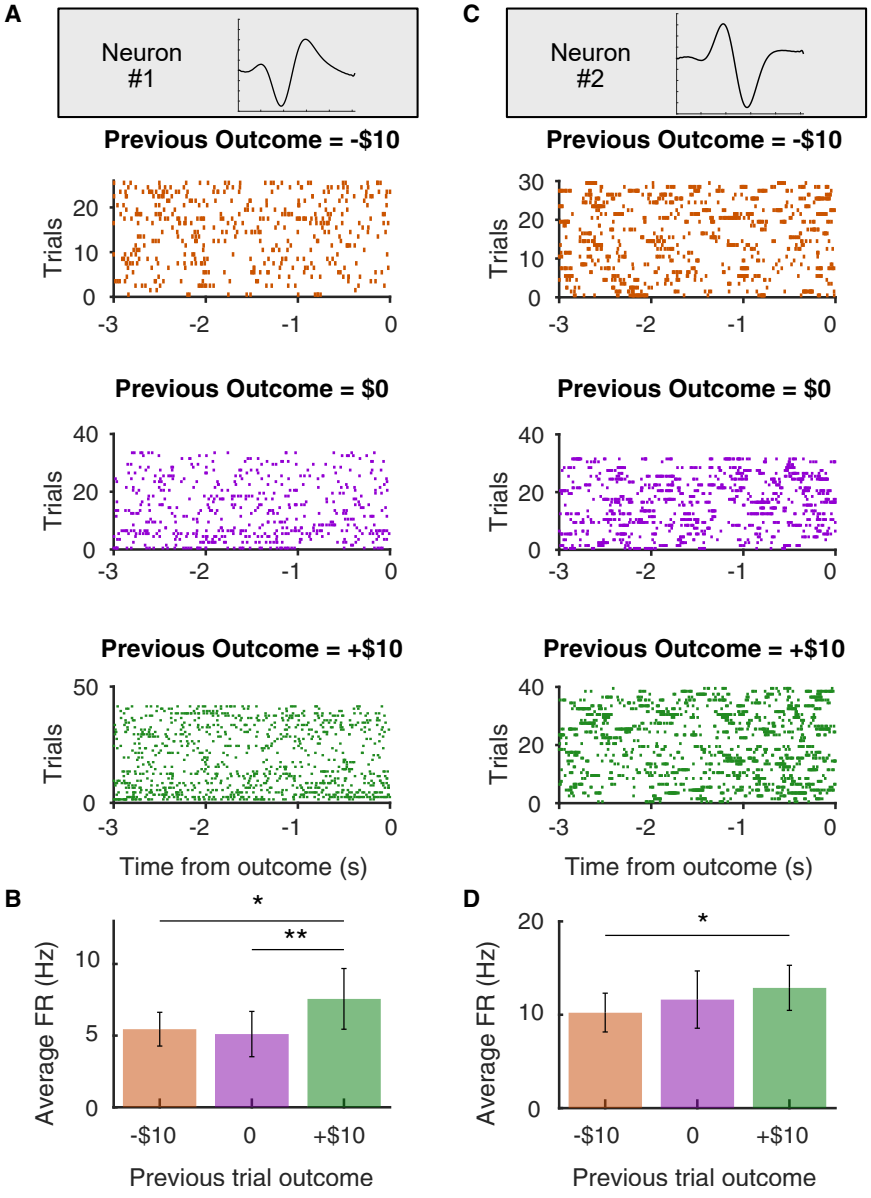



**Figure 3. Sustained firing rate during the expectation period is modulated by the previous reward outcome**

(A) Activity raster plots from a representative neuron (waveforms shown of each neuron), separated according to previous trial outcome: negative (-$10, top), neutral ($0, middle) and positive (+$10), bottom during the expectation period prior between patient choice (t = −3000 ms) and reward reveal outcome (t = 0).

(B) Average firing rate ± standard error of the mean during the expectation period split by previous reward outcome, averaged across trials. The sustained firing rate was significantly greater in trials following positive outcomes than in trials following negative outcomes ($p < 0.05$, paired $t$ test) and neutral outcomes ($p < 0.01$, paired $t$ test).

(C and D) as (A and B), for a second neuron that only showed differences between negative (-$10) and positive (+$10) previous outcomes.

choices under this model were heavily dependent on the previous trial. Overall, the computational modeling results indicate that either an RW model with a high learning rate or a heuristic win-stay/lose-switch model can capture the patients' behavior. Under either of these strategies, choices are strongly dependent on the outcome of the previous trial, consistent with the notion that reward expectations can drive behavior in both gradual learning (i.e., in RL) as well as short-term processes (i.e., win-stay/lose-switch heuristics).

## Firing rate during expectation is modulated by prior outcome

Next, we sought to examine whether individual SN neurons encode reward history information. We analyzed the FR of putative DA neurons ($n$ = 22) during the reward expectation period, after patient choice but before outcome reveal (3 s; Figure 1B). Since our modeling results indicated that patients' choices were strongly determined by prior trial outcomes (under either the RW or heuristic models; Figure 2), we hypothesized that neural signals reflecting reward expectation during this period should depend on previous trial outcomes. Because operating room (OR) time limitations meant that within-block analyses would have resulted in small numbers of trials per block and low statistical power, instead we opted to group the neural activity by previous trial outcome (+10, 0, or −10) independent of block. We observed that neuronal FR during the expectation period was dependent on prior trial outcome, with neurons showing higher FRs in trials following positive trials (+$10) than in negative trials (-$10; two representative examples shown in Figure 3). Some neurons also showed a difference in FR between neutral ($0) and negative (-$10) outcomes, reflecting a gradual increase in expectation period FRs going from negative to neutral to positive outcomes (Figure 3B). To quantitatively examine the relationship at the trial level and to help account for potential confounding factors, we used a generalized linear model (GLM) and linear mixed model (LMM) approach.[26] First, we assessed the relationships among the independent variables by running a correlation matrix analysis between the previous motor action (left or right), previous trial reward identity (win or loss), previous trial reward outcome (+10, 0, or −10), and previous trial RPE. We found that previous trial reward identity and previous trial reward outcome were highly correlated ($r = 0.81$, $p < 0.05$), whereas no significant correlations were observed between the other variables ($p > 0.05$). Before constructing a final model, we ran univariate GLMs to identify which

variables significantly contributed to the expectation period FR. We found that previous trial reward identity and previous trial reward outcome were significantly associated with the expectation period FR (previous reward identity $p$ = 0.0011, t(2240) = 3.2718; previous reward outcome $p$ = 6.694e-8, t(2240) = 5.4174; previous action $p$ = 0.7464, t(2240) = 0.3235; previous RPE $p$ = 0.3398, t(2240) = −0.9547). Given the strong correlation between previous reward identity and previous reward outcome, and the lower AIC of the model including previous reward outcome (146.2 compared to 159.0), the final linear model included previous motor action, previous trial reward outcome, and previous trial RPE as regressors for the expectation period FR. We ran this model for each individual neuron and found that 8/22 neurons showed a significant effect of previous reward outcome on the expectation period FR, while no neurons showed significance for previous action or previous RPE. After running the model on individual neurons, we assessed the group-level effect, accounting for individual differences across neurons by including a random intercept. Among these variables, previous reward outcome was significantly associated with FRs during the expectation period (previous reward outcome $p$ = 6.91e-8, t(2217) = 5.412; previous action $p$ = 0.8901, t(2217) = 0.138; previous RPE $p$ = 0.3264, t(2217) = −0.982) (Table S2), with higher FRs in trials following rewards. To examine whether these effects were specific to the expectation period, or whether they reflect sustained FR modulation across trials, we run a similar GLM during the choice period (after slot machine presentation, but before choice). Although we found a significant effect of previous reward outcome on FR during the choice period ($p$ = 0.0397, t(2217) = 2.058), the effect was markedly smaller than during the expectation period, where modulation was strongest. A similar analysis including an interaction term between previous reward outcome and trial types did not reveal a significant interaction effect ($p$ = 0.12), suggesting that the relationship between previous outcome and FR is consistent across blocks, and that pooling all task trials across does not mask block effects. Finally, we examined whether neuronal activity reflected motor activity by examining the average FR during the decision period, but observed no movement-related modulation at button press (Figure S5A).

## Diverse firing rate responses to feedback in putative DA neurons

Previous studies in rodents[8,27–29] and humans[6] have demonstrated that individual DA neurons encode unexpected reward outcomes through FR modulation. To verify whether this was the case in our dataset, we also examined putative DA responses after reward outcome (feedback epoch) and compared FR following rewarded (i.e., positive outcome trials) vs. unrewarded (i.e., negative outcome trials). In this feedback epoch, we found a subset of SN neurons (7/22, or 31.8%) whose FR was modulated according to the reward outcome (Figure S3). Significance was determined using individual uncorrected t-tests at each time point post feedback. After doing more stringent permutation testing to determine clusters of significance, only 1/22 neurons showed a significant cluster difference in the post-feedback time domain. Interestingly, whereas some showed an increase in FR following unexpected rewards (4/22, or 18.2%), a different set of neurons showed the opposite pattern, with FR decreasing following rewarded trials (3/22, 13.6%). In alignment with our expectation period analysis, we also conducted a GLM analysis during the post-reward outcome epoch (FR during the feedback period), including the current action (left or right), current trial reward outcome (+10, 0, or −10), and current RPE as regressors. None of these variables had a significant effect on FR during the feedback period (reward outcome: $p$ = 0.0559, t(2283) = 1.912; action: $p$ = 0.5068, t(2283) = 0.664; RPE: $p$ = 0.2926, t(2283) = −1.053). Overall, we found limited evidence for reward encoding at the group level, which was especially salient in the mixed block where the value difference between the different potential outcomes was maximal (Figure S4). Therefore, we found some evidence for reward encoding in putative DA neurons, but with a certain amount of variation across individual neuronal responses.

## Group-level encoding of reward history in neural activity

Next, we examined whether the reward history encoding was present at the group level by comparing the FR of all putative DA neurons during the expectation period across previous outcome identities (-$10/$0/+$10). To compensate for the heterogeneity in baseline FRs across neurons, we normalized FRs to the average expectation period FR per neuron across all trials. For each neuron, we averaged the normalized FRs across trial types. We found that FRs of putative DA neurons were higher in trials following positive (+$10) vs. negative (-$10; $p$ = 0.007, Wilcoxon test; zval = 2.7346) and neutral outcomes ($0; $p$ = 0.034, Wilcoxon test; zval = 2.1243; Figures 4A and 4B), but no significant difference between trials following negative and neutral outcomes (-$10 and $0, $p$ = 0.285, Wilcoxon test; zval = 1.0680).

Previous studies have shown a pre-reward ramping up of dopaminergic activity as animals approach reward.[17,30–35] To test for similar ramping dynamics, we carried out a linear regression between FR and time elapsed during the expectation period, where a significant positive relationship would indicate the presence of ramping. However, we did not observe such an association (linear correlation coefficient = 0.0825, $p$ = 0.753), suggesting that ramping dynamics were not present in our dataset.

## Group-level encoding of reward expectation in reaction times

Finally, motivated by previous studies and models linking tonic DA levels to response vigor,[36] we examined whether differences in prior outcomes also affected RTs. We found that RTs were faster following trials with positive outcomes compared to after neutral/negative outcomes (+$10 median = 0.916 s vs. -$10/$0 median = 1.017 s, $p$ = 0.0175, Wilcoxon rank-sum test, Figure 4C). Beyond this behavioral effect, we sought to assess whether a direct association existed between putative dopaminergic neuron spiking and subsequent RTs. A mixed model analysis revealed a significant association between expectation period FRs and subsequent RTs (estimate: −531.87, $p$ = 0.0458), with block type included as a fixed effect and patient ID as a random effect to account for inter-patient variability. When the fixed effect for block type was removed, the $p$-value increased to 0.0529, suggesting that the association is modest. We note that the prior trial outcome was not included as a

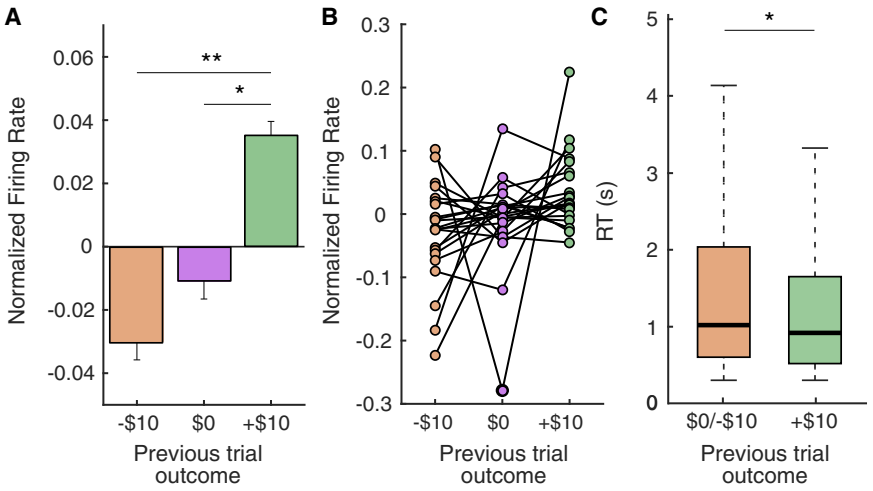



**Figure 4. Putative DA neurons encode previous reward outcome during the expectation period**

(A) Average normalized firing rate across all putative DA neurons during the reward expectation period (3 s post-choice, prior to reward outcome reveal), separated according to previous trial outcome: negative (-$10), neutral ($0), or positive (+$10). FR of putative DA neurons is higher for previous positive outcomes than either previous neutral ($p < 0.05$, Wilcoxon rank sum) or previous negative outcomes ($p < 0.01$, Wilcoxon rank sum).

(B) Same data as (A) but showing individual putative DA neurons.

(C) Patient reaction times (RTs) varied depending on the previous trial outcome. RTs were faster following a positive, compared to neutral or negative outcomes (+$10 median = 0.916 s vs. -$10/$0 median = 1.017 s, $p < 0.05$, Wilcoxon rank sum).

covariate in this model, and therefore, residual influences of prior reward cannot be fully excluded. Nonetheless, our findings suggest that trial-by-trial variations in expectation period FRs may influence subsequent RTs, with higher FRs associated with faster responses. At a broader level, we observed that the sustained FRs of SN neurons were modulated during the reward expectation period, with higher FRs during the expectation period following positive outcomes. In addition to these neural effects, previous reward outcomes also had a behavioral correlate, as participants exhibited faster responses following positive outcomes.

## DISCUSSION

The ability to compute differences between expected and obtained rewards is fundamental for adaptive decision-making in uncertain environments. This computation critically depends on the previous history of reward, which informs reward expectations. Despite the known role of the dopaminergic system in encoding RPEs, how these processes are represented in the human brain remains elusive due to difficulties recording from subcortical dopaminergic regions. Here, building on previous studies using intraoperative single unit recordings[6,7] and fast-scan cyclic voltammetry[19,20] we aimed to characterize how spiking activity of putative DA neurons in the human SN reflects recent reward outcomes during a risky decision-making task.

### Choice behavior is influenced by previous reward outcomes

Computational modeling revealed that patients' behavior could be captured by a win-stay/lose-shift heuristic or a high-learning-rate RW model, both emphasizing a strong influence of the outcome of the immediately prior trial. Conceptually, the previous trial outcome can be thought of as a reward expectation signal that does not necessarily rely on incremental learning processes (i.e., reinforcement learning processes). Rather, previous positive outcomes can directly lead to optimistic expectations, resulting in a higher proportion of "stay" choices, while negative outcomes lead to pessimistic expectations and strategic switches. Previous trial outcomes had a neurophysiological correlate, with the FRs of putative DA neurons modulated by the reward outcome of the preceding trial: FRs increased following positive outcomes (+$10) compared to neutral ($0) and negative (-$10) outcomes (Figure 3). Thus, previous trial outcomes exerted a lingering influence on nigral activity during the expectation period, with FRs scaling according to the outcome's valence. These results are broadly consistent with previous observations in animal models that the tonic firing of DA neurons tracks values, with sustained firing increases following signaled value increases[17,32,37] and tonic activity encoding the net reward rate.[36] Recent evidence from animal models shows that DA transients in the striatum track reward rates, but potentially not in VTA DA responses,[32] whereas our results, in contrast, suggest that human SN neurons can signal reward expectation rates.

In addition to value encoding, tonic DA activity may enhance motivation and facilitate faster motor responses in high-reward contexts,[38] consistent with our finding that patient RTs were significantly faster following positive outcomes (Figure 4C). Additionally, we observed a negative trend between FRs of putative DA neurons during the expectation period and subsequent RTs, suggesting that higher DA activity is associated with faster reactions. This aligns with previous human PET[39] and behavioral experiments[40,41] showing that high expectations modulate the DA system and enhance vigor,[36,42] resulting in faster RTs. Alternatively, these effects may be explained by post-error slowing after negative outcomes.[43,44]

### Variability in post-reward responses of putative DA neurons

Our study focuses on the pre-reward period, whereas most human electrophysiology studies have focused on the post-reward period. While we observed responses resembling RPEs in some dopaminergic neurons (Figures S3 and S4), post-reward responses were less robust than previous findings,[6] with only a minority of neurons showing such effects. Uncorrected analyses suggested heterogeneity across a few neurons (with 4 neurons showing post-outcome negative encoding, in which FRs were higher for losses, and 3 neurons showing positive encoding;

Figure S3). In addition, only one neuron survived more stringent permutation tests, further demonstrating weak post-reward responses. As a consequence, we have limited ability to test whether significant heterogeneity in neuronal encoding, which has been previously shown in animal models[16,45] also exists in human nigra neurons. For example, dopaminergic neurons in rodent SN and VTA are capable of reflecting positive, negative RPEs or unsigned salience,[45–49] with differences in the representation of RPE in SN and VTA[50]; our results show weak evidence in support of this heterogeneity in the human nigra. Future experiments focused on characterizing the heterogeneity in response of putative DA post-outcome responses in the human SN will constitute a promising future direction. Finally, we examined whether neurons with weak post-reward encoding also reflected previous reward history, by testing whether neurons that reflected encoded both types of signal were more frequent than expected, but found no evidence to suggest this (chi-square test, $p = 0.665$). Therefore, we found no evidence to suggest that there is selective mixing of previous reward history and weak post-reward encoding. Future experiments focused on characterizing the heterogeneity in response of putative DA post-outcome responses in the human SN will constitute a promising future direction.

### From reward history to outcome

According to simple RPE models, RPEs arise from the arithmetic combination of expectation and reward information[1,51] but the nature of their neural representations remains to be fully determined. Some models assume independent representation of both quantities converging at DA neurons.[1,51–54] Alternatively, reward and reward expectations may be computed redundantly in several different regions,[16] with other expectation and reward signals combined in upstream input regions.[55–57] Within this framework, our data suggest that the sustained FRs of putative DA neurons during the pre-feedback period are shaped by reward history. Such signals may align with, but are not limited to, processes related to reward expectations, emphasizing that our definition captures a short-horizon predictive component derived from reward history. More specifically, the encoding of reward history during the pre-feedback period may provide the baseline against which subsequent outcomes are evaluated. In other words, by shaping the predicted value of the current choice, reward history signals could determine the reference point that is subtracted from actual reward at the time of feedback. This mechanism would naturally contribute to the computation of RPEs, even if such signals are not identical to formal model-derived expected values.

In summary, combining intraoperative SUA recordings with computational modeling, we show that sustained FR modulation in human SN neurons reflects prior trial outcomes, a signal that could be related to reward history signals. Our findings provide preliminary insights into the role of individual human nigra neurons, representing recent outcomes, and add evidence to the suggested link between dopaminergic dynamics and motivational vigor,[58,59] enhancing our understanding of reward learning and motivation in the human brain with implications for disorders involving dopaminergic dysfunction.

### Limitations of the study

First, our dataset comprised 27 neurons recorded across 13 sessions in 11 patients with PD (average = 2.45 ± 0.82 neurons per patient). This small sample size reduces statistical power and limits our ability to characterize population-level coding, individual neuronal coding or between-subject variability. As a consequence, given the limited power to make inferences on individual neuron coding (Figure S3), we instead focus on population-level results, which were statistically significant despite these dataset limitations (Figure 4). Second, all recordings were obtained from a PD cohort undergoing DBS surgery; while previous studies using similar intraoperative methods have yielded valuable insights into human dopaminergic function, caution is warranted when generalizing our findings to healthy individuals, as chronic dopaminergic degeneration in PD may alter neuronal physiology. Moreover, we acknowledge the difficulty of including an appropriate control group for comparison in intraoperative human recordings, which is not possible in healthy participants and further constrains the interpretability of our results. Third, although we successfully selected lower FR neurons, resulting in a high proportion of putative DA neurons ($n$ = 22/27) compared to presumed GABAergic neurons ($n$ = 5/27), classification relied primarily on FR and spike waveform width. We chose to focus exclusively on these two metrics for simplicity and for consistency with previous human studies.[6,7] Due to our experimental limitations, neurochemical confirmation was not possible in this context, limiting our ability to directly estimate individual cells' signaling profiles. Fourth, the behavioral task was conducted in a short period of time intraoperatively, which may have made it difficult for patients to engage in incremental learning processes. Fifth, although we chose to use Osort for consistency with previous human single-unit studies,[60] it is likely that other methods, such as Kilosort[61] or UMAP-based sorting[62] would also be effective at spike sorting human data; future studies comparing these methods may be valuable. Finally, limitations in signal-to-noise ratio resulted in insufficient statistical power for separate analyses per block, precluding finer-grained within-block analyses.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to the lead contact, Dr. Ignacio Saez (ignacio.saez@mssm.edu).

### Materials availability

This study did not generate any material reagents.

### Data and code availability

- The datasets generated during the current study are available from the corresponding author on reasonable request.
- All code necessary for data cleaning and analysis is available in our associated repository at https://github.com/zarghonai/single_unit.
- Any additional data and information needed to reanalyze the data are available upon reasonable request from the lead contact.

## ACKNOWLEDGMENTS

We are grateful to all patients who participated in this study. We would like to thank Juri Minxha for assisting in the use of OSort and the surgical team at Mount Sinai West for support during data collection. We would also like to

thank Alice Hashemi and Kimia Ziafat for consenting the patients for the research. We are grateful to Naoshige Uchida and members of the Saez and Gu laboratory for their comments on the manuscript. A.K. is supported by JSPS Overseas Research Fellowships. X.G. is supported by the National Institute on Drug Abuse [grant nos.: R01DA043695 and R21DA049243] and National Institute of Mental Health [grant nos: R21MH120789, R01MH122611, R01MH123069, and R01MH124115]. I.S. is supported by the National Institute of Mental Health [grant nos.: R01MH124763 and R01MH104606] and the National Science Foundation [project no.: 2319580].

## AUTHOR CONTRIBUTIONS

I.S., X.G., and B.H.K. designed the study. Z.I., L.N.M., A.N.D., M.H., A.M., and B.H.K. were involved in data collection. K.K. and X.G. designed the task. Z.I., S.E.Q., A.K., and A.N.D. carried out neural data analyses; Z.I., S.E.Q., and A.K. carried out behavioral analyses. I.S. and X.G. supervised all analyses. B.H.K. and A.W.C. supervised surgical aspects of the project. Z.I., A.K., and I.S. wrote the original draft of the paper. All authors participated in editing and approved the final document.

## DECLARATION OF INTERESTS

B.H.K. is a consultant for Medtronic, Abbott Neuromodulation, and Turing Medical.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Deep brain stimulation surgery
  - Behavioral paradigm
  - Computational modeling of behavior
  - Model implementation and inference
  - Electrophysiological data collection
  - Linear mixed model analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.115458.

Received: June 3, 2025
Revised: October 13, 2025
Accepted: March 19, 2026
Published: March 21, 2026

## REFERENCES

1. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
2. Niv, Y. (2009). Reinforcement learning in the brain. J. Math. Psychol. *53*, 139–154. https://doi.org/10.1016/j.jmp.2008.12.005.
3. Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. Nature *525*, 243–246. https://doi.org/10.1038/nature14855.
4. McClure, S.M., Berns, G.S., and Montague, P.R. (2003). Temporal prediction errors in a passive learning task activate human striatum. Neuron *38*, 339–346. https://doi.org/10.1016/s0896-6273(03)00154-5.
5. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337. https://doi.org/10.1016/s0896-6273(03)00169-7.
6. Zaghloul, K.A., Blanco, J.A., Weidemann, C.T., McGill, K., Jaggi, J.L., Baltuch, G.H., and Kahana, M.J. (2009). Human substantia nigra neurons encode unexpected financial rewards. Science *323*, 1496–1499. https://doi.org/10.1126/science.1167342.
7. Ramayya, A.G., Misra, A., Baltuch, G.H., and Kahana, M.J. (2014). Microstimulation of the human substantia nigra alters reinforcement learning. J. Neurosci. *34*, 6887–6895. https://doi.org/10.1523/jneurosci.5445-13.2014.
8. Schultz, W. (2016). Dopamine reward prediction error coding. Dialogues Clin. Neurosci. *18*, 23–32. https://doi.org/10.31887/DCNS.2016.18.1/wschultz.
9. Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. J. Neurosci. *12*, 4595–4610. https://doi.org/10.1523/jneurosci.12-12-04595.1992.
10. Iturra-Mena, A.M., Kangas, B.D., Luc, O.T., Potter, D., and Pizzagalli, D.A. (2023). Electrophysiological signatures of reward learning in the rodent touchscreen-based Probabilistic Reward Task. Neuropsychopharmacology *48*, 700–709. https://doi.org/10.1038/s41386-023-01532-4.
11. Simon, N.W., and Moghaddam, B. (2015). Neural processing of reward in adolescent rodents. Dev. Cogn. Neurosci. *11*, 145–154. https://doi.org/10.1016/j.dcn.2014.11.001.
12. Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. Annu. Rev. Neurosci. *40*, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109.
13. Hikosaka, O., Sakamoto, M., and Usui, S. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. J. Neurophysiol. *61*, 814–832. https://doi.org/10.1152/jn.1989.61.4.814.
14. Schultz, W., Tremblay, L., and Hollerman, J.R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb. Cortex *10*, 272–284. https://doi.org/10.1093/cercor/10.3.272.
15. Farries, M.A., Faust, T.W., Mohebi, A., and Berke, J.D. (2023). Selective encoding of reward predictions and prediction errors by globus pallidus subpopulations. Curr. Biol. *33*, 4124–4135.e5. https://doi.org/10.1016/j.cub.2023.08.042.
16. Tian, J., Huang, R., Cohen, J.Y., Osakada, F., Kobak, D., Machens, C.K., Callaway, E.M., Uchida, N., and Watabe-Uchida, M. (2016). Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons. Neuron *91*, 1374–1389. https://doi.org/10.1016/j.neuron.2016.08.018.
17. Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat. Neurosci. *19*, 117–126. https://doi.org/10.1038/nn.4173.
18. Collins, A.L., Greenfield, V.Y., Bye, J.K., Linker, K.E., Wang, A.S., and Wassum, K.M. (2016). Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. Sci. Rep. *6*, 20231. https://doi.org/10.1038/srep20231.
19. Kishida, K.T., Saez, I., Lohrenz, T., Witcher, M.R., Laxton, A.W., Tatter, S.B., White, J.P., Ellis, T.L., Phillips, P.E.M., and Montague, P.R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. Proc. Natl. Acad. Sci. USA *113*, 200–205. https://doi.org/10.1073/pnas.1513619112.
20. Moran, R.J., Kishida, K.T., Lohrenz, T., Saez, I., Laxton, A.W., Witcher, M.R., Tatter, S.B., Ellis, T.L., Phillips, P.E., Dayan, P., and Montague, P.R. (2018). The Protective Action Encoding of Serotonin Transients in the Human Brain. Neuropsychopharmacology *43*, 1425–1435. https://doi.org/10.1038/npp.2017.304.
21. Sands, L.P., Jiang, A., Liebenow, B., DiMarco, E., Laxton, A.W., Tatter, S.B., Montague, P.R., and Kishida, K.T. (2023). Subsecond fluctuations in extracellular dopamine encode reward and punishment prediction

errors in humans. Sci. Adv. *9*, eadi4927. https://doi.org/10.1126/sciadv.adi4927.

22. Ramayya, A.G., Zaghloul, K.A., Weidemann, C.T., Baltuch, G.H., and Kahana, M.J. (2014). Electrophysiological evidence for functionally distinct neuronal populations in the human substantia nigra. Front. Hum. Neurosci. *8*, 655. https://doi.org/10.3389/fnhum.2014.00655.

23. Yau, J.O.-Y., and McNally, G.P. (2023). The Rescorla-Wagner model, prediction error, and fear learning. Neurobiol. Learn. Mem. *203*, 107799. https://doi.org/10.1016/j.nlm.2023.107799.

24. Siegel, S., and Allan, L.G. (1996). The widespread influence of the Rescorla-Wagner model. Psychon. Bull. Rev. *3*, 314–321. https://doi.org/10.3758/BF03210755.

25. Miller, R.R., Barnet, R.C., and Grahame, N.J. (1995). Assessment of the Rescorla-Wagner model. Psychol. Bull. *117*, 363–386. https://doi.org/10.1037/0033-2909.117.3.363.

26. Arnold, K.F., Davies, V., de Kamps, M., Tennant, P.W.G., Mbotwa, J., and Gilthorpe, M.S. (2021). Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. Int. J. Epidemiol. *49*, 2074–2082. https://doi.org/10.1093/ije/dyaa049.

27. Hollerman, J.R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. *1*, 304–309. https://doi.org/10.1038/1124.

28. Sturman, D.A., and Moghaddam, B. (2011). Reduced neuronal inhibition and coordination of adolescent prefrontal cortex during motivated behavior. J. Neurosci. *31*, 1471–1478. https://doi.org/10.1523/jneurosci.4210-10.2011.

29. Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., and Schoenbaum, G. (2009). The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron *62*, 269–280. https://doi.org/10.1016/j.neuron.2009.03.005.

30. Roitman, M.F., Wheeler, R.A., and Carelli, R.M. (2005). Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. Neuron *45*, 587–597. https://doi.org/10.1016/j.neuron.2004.12.055.

31. Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature *500*, 575–579. https://doi.org/10.1038/nature12475.

32. Mohebi, A., Pettibone, J.R., Hamid, A.A., Wong, J.M.T., Vinson, L.T., Patriarchi, T., Tian, L., Kennedy, R.T., and Berke, J.D. (2019). Dissociable dopamine dynamics for learning and motivation. Nature *570*, 65–70. https://doi.org/10.1038/s41586-019-1235-y.

33. Kim, H.R., Malik, A.N., Mikhael, J.G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S.J., and Uchida, N. (2020). A Unified Framework for Dopamine Signals across Timescales. Cell *183*, 1600–1616.e25. https://doi.org/10.1016/j.cell.2020.11.013.

34. Farrell, K., Lak, A., and Saleem, A.B. (2022). Midbrain dopamine neurons signal phasic and ramping reward prediction error during goal-directed navigation. Cell Rep. *41*, 111470. https://doi.org/10.1016/j.celrep.2022.111470.

35. Sarno, S., Beirán, M., Falcó-Roget, J., Diaz-deLeon, G., Rossi-Pool, R., Romo, R., and Parga, N. (2022). Dopamine firing plays a dual role in coding reward prediction errors and signaling motivation in a working memory task. Proc. Natl. Acad. Sci. USA *119*, e2113311119. https://doi.org/10.1073/pnas.2113311119.

36. Niv, Y. (2007). Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? Ann. N. Y. Acad. Sci. *1104*, 357–376. https://doi.org/10.1196/annals.1390.018.

37. Wang, Y., Toyoshima, O., Kunimatsu, J., Yamada, H., and Matsumoto, M. (2021). Tonic firing mode of midbrain dopamine neurons continuously tracks reward values changing moment-by-moment. eLife *10*, e63166. https://doi.org/10.7554/eLife.63166.

38. Cagniard, B., Balsam, P.D., Brunner, D., and Zhuang, X. (2006). Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. Neuropsychopharmacology *31*, 1362–1370. https://doi.org/10.1038/sj.npp.1300966.

39. Ando, S., Fujimoto, T., Sudo, M., Watanuki, S., Hiraoka, K., Takeda, K., Takagi, Y., Kitajima, D., Mochizuki, K., Matsuura, K., et al. (2024). The neuromodulatory role of dopamine in improved reaction time by acute cardiovascular exercise. J. Physiol. *602*, 461–484. https://doi.org/10.1113/jp285173.

40. Sedaghat-Nejad, E., Pi, J.S., Hage, P., Fakharian, M.A., and Shadmehr, R. (2022). Synchronous spiking of cerebellar Purkinje cells during control of movements. Proc. Natl. Acad. Sci. USA *119*, e2118954119. https://doi.org/10.1073/pnas.2118954119.

41. Guitart-Masip, M., Beierholm, U.R., Dolan, R., Duzel, E., and Dayan, P. (2011). Vigor in the face of fluctuating rates of reward: an experimental examination. J. Cogn. Neurosci. *23*, 3933–3938. https://doi.org/10.1162/jocn_a_00090.

42. Berke, J.D. (2018). What does dopamine mean? Nat. Neurosci. *21*, 787–793. https://doi.org/10.1038/s41593-018-0152-y.

43. Wang, L., Tang, D., Zhao, Y., Hitchman, G., Wu, S., Tan, J., and Chen, A. (2015). Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing. Sci. Rep. *5*, 8708. https://doi.org/10.1038/srep08708.

44. Rabbitt, P.M. (1966). Errors and error correction in choice-response tasks. J. Exp. Psychol. *71*, 264–272. https://doi.org/10.1037/h0022853.

45. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature *577*, 671–675. https://doi.org/10.1038/s41586-019-1924-6.

46. Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898–1902. https://doi.org/10.1126/science.1077349.

47. Lammel, S., Lim, B.K., and Malenka, R.C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. Neuropharmacology *76*, 351–359. https://doi.org/10.1016/j.neuropharm.2013.03.019.

48. Joshua, M., Adler, A., Mitelman, R., Vaadia, E., and Bergman, H. (2008). Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials. J. Neurosci. *28*, 11673–11684. https://doi.org/10.1523/jneurosci.3839-08.2008.

49. Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R., and Deisseroth, K. (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell *162*, 635–647. https://doi.org/10.1016/j.cell.2015.07.014.

50. Keiflin, R., Pribut, H.J., Shah, N.B., and Janak, P.H. (2019). Ventral Tegmental Dopamine Neurons Participate in Reward Identity Predictions. Curr. Biol. *29*, 93–103.e3. https://doi.org/10.1016/j.cub.2018.11.050.

51. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (A Bradford Book).

52. Kawato, M., and Samejima, K. (2007). Efficient reinforcement learning: computational theories, neuroscience and robotics. Curr. Opin. Neurobiol. *17*, 205–212. https://doi.org/10.1016/j.conb.2007.03.004.

53. Morita, K., Morishima, M., Sakai, K., and Kawaguchi, Y. (2013). Dopaminergic control of motivation and reinforcement learning: a closed-circuit account for reward-oriented behavior. J. Neurosci. *33*, 8866–8890. https://doi.org/10.1523/jneurosci.4614-12.2013.

54. James, C., Houk, J.L.D., and Beiser, D.G. (1994). Models of Information Processing in the Basal Ganglia (The MIT Press).

55. Contreras-Vidal, J.L., and Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. J. Comput. Neurosci. *6*, 191–214. https://doi.org/10.1023/a:1008862904946.

56. Hazy, T.E., Frank, M.J., and O'Reilly, R.C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. Neurosci. Biobehav. Rev. *34*, 701–720. https://doi.org/10.1016/j.neubiorev.2009.11.019.

57. Tan, C.O., and Bullock, D. (2008). A local circuit model of learned striatal and dopamine cell responses under probabilistic schedules of reward. J. Neurosci. *28*, 10062–10074. https://doi.org/10.1523/jneurosci.0259-08.2008.

58. Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., and Dayan, P. (2013). Dopamine Modulates Reward-Related Vigor. Neuropsychopharmacology *38*, 1495–1503. https://doi.org/10.1038/npp.2013.48.

59. Hofmans, L., Westbrook, A., van den Bosch, R., Booij, J., Verkes, R.-J., and Cools, R. (2022). Effects of average reward rate on vigor as a function of individual variation in striatal dopamine. Psychopharmacology *239*, 465–478. https://doi.org/10.1007/s00213-021-06017-0.

60. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2006). Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J. Neurosci. Methods *154*, 204–224. https://doi.org/10.1016/j.jneumeth.2005.12.033.

61. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M., and Kenneth D, H. (2016). Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. Preprint at bioRxiv. https://doi.org/10.1101/061481.

62. Sedaghat-Nejad, E., Fakharian, M.A., Pi, J., Hage, P., Kojima, Y., Soetedjo, R., Ohmae, S., Medina, J.F., and Shadmehr, R. (2021). P-sort: an open-source software for cerebellar neurophysiology. J. Neurophysiol. *126*, 1055–1075. https://doi.org/10.1152/jn.00172.2021.

63. Ramayya, A.G., Pedisich, I., Levy, D., Lyalenko, A., Wanda, P., Rizzuto, D., Baltuch, G.H., and Kahana, M.J. (2017). Proximity of Substantia Nigra Microstimulation to Putative GABAergic Neurons Predicts Modulation of Human Reinforcement Learning. Front. Hum. Neurosci. *11*, 200. https://doi.org/10.3389/fnhum.2017.00200.

64. Rescorla, R.A., and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning: current research and theory (Appleton-Century-Crofts).

65. Wilson, R.C., and Collins, A.G. (2019). Ten simple rules for the computational modeling of behavioral data. eLife *8*, e49547. https://doi.org/10.7554/eLife.49547.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| Original Code | GitHub | https://github.com/zarghonai/single_unit |
| Software and algorithms | | |
| MATLAB | MATLAB | https://www.mathworks.com/products/matlab.html |
| OSort | Rutishauser et al. | https://www.urut.ch/new/serendipity/index.php?/pages/osort.html |

### EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Research protocol was administered in accordance with the Icahn School of Medicine at Mount Sinai Institutional Review Board (Study 13-00415). Participants age and sex are listed in Table S1. There were 13 sessions from a total of 11 participants (n=11; 7 male, 4 female, average age = 65.71 ± 7.69 years). Due to the limited sample size, influence of sex and gender was not examined on the results.

### METHOD DETAILS

#### Deep brain stimulation surgery

We used intraoperative microelectrode recordings from patients undergoing DBS surgery for Parkinson's disease. Patients were consented for research prior to their surgery date. Patients were off their medications at least 9 hours prior to the surgery as per surgical protocol. The instructions of the task were explained during the consent and again before playing the task. The surgery proceeded as per clinical routine and research was conducted when the patients were woken up for clinical testing during DBS targeting. A tungsten microelectrode with power-assisted Microdrive was lowered through the brain following the tractography previously determined by the clinical team.[63] The surgeon (B.K.) lowered the electrode into the SNr after determining the ventral border of the STN and then listened for unit firing to confirm the location of the electrode (Figure 1A). The recordings were done from the dorsal substantia nigra. Once the surgeon was satisfied that stable spikes could be detected, the electrode was stationary for the duration of the experiment. To maximize our chances of recording from dopaminergic neurons which have lower FRs, we actively sought quieter neurons with lower spike frequencies, versus highly active neurons which were likely to be identified as putative GABAergic neurons. The patient was woken up during the surgery once the electrode was implanted for clinical testing, at which point they performed the task. Research time during the surgery was limited to 15 minutes or less (average = 10.30±2.48 min).

#### Behavioral paradigm

Patients played a two-armed bandit task in which they had to press a button to choose between two slot machines with initially unknown reward probabilities. Patients were taught the layout of the task and had a chance to practice in the days prior to their surgery and then again right before starting the game during surgery. Patients were instructed to maximize their reward by finding and choosing the machine that had the higher probability of resulting in the better outcome. The task had three blocks with 35 trials per block: a reward block, a punishment block, and a mixed block, whose order was randomized across patients. Each block had different win and loss outcomes; the reward block had positive (+$10) or neutral (+$0) outcomes; the punishment block had negative (-$10) and neutral (-$0) outcomes, and the mixed block had positive (+$10) and negative (-$10) outcomes. The two machines had different probabilities of resulting in the better outcome. One had a high probability (80%/20%) and the other had a low probability (20%/80%) (Figure 1B). These probabilities switched twice throughout each block on trials 12/13 and trials 24/25 depending on the reversal condition. The location of the machines stayed the same throughout the block while though the better machine switched between left and right. The task was balanced between which machine was the better machine throughout the entire task. The patients were made aware of this during the instructions to the task. Patients were taught the layout of the task in the days prior to their surgery and then again right before starting the game during the surgery. Patients were told the objective of the game was to maximize their reward by finding and choosing the machine that had the higher probability of giving the winning amount. The task took approximately 10 minutes on average (mean = 10.30±2.48 minutes), and patients were not compensated for participating in research as per IRB protocol. Only patients who completed the full task were considered for analysis. For reaction times analyses, we measured the time between trial presentation and choice and excluded trials with response times shorter than 300 milliseconds and longer than 12 seconds which are likely due to patient error.

## Computational modeling of behavior

For each patient, we first calculated model-free metrics of behavior: number of wins and losses in each block as well as the number of times the machine with the higher reward probability was picked regardless of whether it actually resulted in the better outcome.

To determine which learning method our patients were using to make their decisions we carried out model comparison in MATLAB. We evaluated a random model in which patients chose randomly between the two options, a heuristic win/stay-lose/shift model, and a Rescorla-Wagner model. We first validated the models testing them with simulated data for each model based on the parameters of the task. For the random model, actions were selected randomly whereas with the heuristic model actions were simulated based on the outcome produced by the previous action (either a win or loss). In the Rescorla-Wagner model, the agent is expected to use the outcomes from previous trials to determine the values of each slot machine and act according to those expected values. The parameters examined in the Rescorla-Wagner model are the learning rate ($\alpha$) and the inverse temperature ($\beta$) which signifies the level of predictability of choice.[64] The value update function can be written as:

$$Q_{t+1}^{k} = Q_{t}^{k} + \alpha\left(r_t - Q_{t}^{k}\right),$$

where $\alpha$ indicates the learning rate, which captures the extent to which the prediction error, updates the cached value, ranging from $\alpha = 0$ for a non-learner agent to $\alpha = 1$for maximal learning from the prediction error. The RPE is captured by the term $\left(r_t - Q_t^k\right)$, indicating the difference between the expected ($Q_t^k$) and the obtained ($r_t$) reward. Therefore, the higher the learning rate, the faster an agent updates their values and takes more recent outcomes into account, whereas a lower learning rate means the agent updates their values at a slower pace and looks at the trend of outcomes beyond the most recent one. The mapping of values to choices is achieved through the following SoftMax equation:

$$p_t^k = \frac{e^{\theta Q_t^k}}{\sum_{i=1}^{k} e^{\theta Q_t^i}},$$

where $\theta$ is the 'inverse temperature' parameter that controls the level of stochasticity in the choice, ranging from $\theta = 0$ for random responding to $\theta = \infty$ for deterministically choosing the highest value option. The learning rate is dependent on both the unconditional and conditional stimuli.[25]

Following data simulation, we did parameter recovery for the Rescorla-Wagner model. Parameter recovery involves using the simulated data with known parameters and fitting the model with the simulated data to recover the parameters back again. The recovered parameters are then compared to the initial parameters to check for correlation. If there is a high correlation, that confirms that the model parameters are identifiable from the task.[65] Finally, we compared the models using a confusion matrix which allows us to see if the data simulated using a certain model is best fit only by that model (Figure S2). If data is better fit by another model that could mean that model is not identifiable with the data and the experimental parameters.[65] We followed these steps first for the entire task without splitting by block followed by splitting by block. We used the outcome to these results to first check if the task was suited for a RL model or if a more complex learning model would be required to model this task. Secondly, by estimating the best model for each of our patients' behaviors, we can confirm that our patients understood the objective of the task and were making informed decisions as opposed to selecting randomly.

## Model implementation and inference

We implemented three candidate choice models in MATLAB and fitted them by maximum likelihood using constrained optimization (fmincon). Random-choice model (M1): included a single parameter b in [0,1], the probability of choosing option 1 (option 2 chosen with probability 1−b). Rewards did not influence choices. The negative log-likelihood (NLL) was $-\Sigma$ log p(a_t). Win–stay/lose–shift model (WSLS; M2): used a symmetric lapse parameter $\varepsilon$ in [0,1]. On the first trial, choices were uniform [0.5, 0.5]. After a rewarded trial (r_{t−1} =1), the model stayed with the previous action with probability $1-\varepsilon/2$ and switched with probability $\varepsilon/2$. After an unrewarded trial, it switched with probability $1-\varepsilon/2$ (stayed with probability $\varepsilon/2$). The NLL was computed from the trial-wise choice probabilities.

$$\mathrm{P(repeat\ choice\ at\ trial\ t)} = \{1 \text{ if reward\_}(t-1) > 0;\ \ 0 \text{ if reward\_}(t-1) \le 0\}$$

Rescorla–Wagner model (RW; M3): estimated a learning rate $\alpha$ in [0,1] and an inverse temperature $\beta$ in [0,50]. Q-values were initialized to Q1 = Q2 = 0.5. On each trial, choice probabilities followed a softmax:

$$\mathrm{p(a\_t = i) = \exp(\beta * Q\_i) \,/\, \Sigma\_j \exp(\beta * Q\_j)}.$$

Values were updated via:

$$\mathrm{\delta\_t = r\_t - Q\_\{a\_t\}, Q\_\{a\_t\} \leftarrow Q\_\{a\_t\} + \alpha * \delta\_t}.$$

We minimized the NLL returned by the likelihood functions for each model. Initial values followed the code defaults (e.g., RW: $\alpha \sim$ Uniform(0,1), $\beta \sim$ Exponential(mean = 1)). Model comparison used the Bayesian Information Criterion (BIC = kln(n) − 2LL, with k the number of free parameters and n the number of trials).

### Electrophysiological data collection

Neural data was recorded from an Alpha Omega machine (Alpha Omega, Nazareth Illit, Israel) routinely used for microelectrode recordings in our DBS cases. Neural data was digitized at a rate of 44000 Hz and filtered between 300-9000 Hz to isolate high-frequency activity reflecting neuronal spiking. Filtered data was analyzed with the OSort spike sorting software.[60] Briefly, spikes were detected if the bandpass filtered signal crossed a threshold determined as five times the standard deviation of the average amplitude of the signal. Once a new spike was detected, the multidimensional distance from the spike to clusters of similar spikes already formed was calculated. If the distance was smaller than a threshold calculated as the square of average standard deviation of the filtered signal calculated along a sliding window times the number of datapoints in a single waveform, then it was added to an existing cluster, but if it was larger than the threshold a new cluster was formed.[60] The clusters were then manually examined for artifactual waveforms, or clusters that were artificially split and merged if necessary (Figure 1C). Clusters were also determined to be artifactual if their inter-spike-interval was within 3 ms greater than 5% of the time. To determine if clusters were artificially split, we analyzed the projection tests outputted by OSort. The projection test compares each cluster against each other by looking at the distance between each cluster and examining if the clusters are far apart enough to consider them two separate clusters.

To distinguish between dopamine neurons and GABAergic interneurons, we calculated the average FR of individual neurons. Putative DA neurons were classified as having an average FR less than 15 Hz, and putative GABAergic interneurons were classified as having an average FR greater than 15 Hz.[22] Neurons with an average FR less than 0.5 Hz were discarded. Additionally, we looked at the waveform width in milliseconds of each neuron which was defined as time from baseline to baseline of the waveform. We classified neurons with a FR below 15 Hz and a waveform width >0.8 ms as putative DA neurons and those with greater than 15 Hz FR and/or a waveform width <0.8 ms as putative GABAergic interneurons.[22]

The neural recordings were epoched time locked to the events in the game. We focused on a range of −3000 ms–1000 ms around outcome reveal. That time frame was chosen because the time in between choice and outcome reveal was 3000 ms and then the outcome stayed on the screen for 1000 ms. Raster plots showing timing of individual spikes were generated for each neuron per block. We then computed the z-scored FR for the average response curve pre, and post outcome reveal for win and loss trials per block. Z-scoring was done by normalizing the FR against the average expectation period FR for the neuron by subtracting the average expectation period FR for the neuron and dividing by the standard deviation of the expectation period FR.

### Linear mixed model analyses

We used linear mixed model (LMM) analyses to examine the relationship between neuronal firing rates and behavioral and computational regressors of interest, while accounting for inter-individual differences. For examining the potential behavioral influences on FR, our LMM included previous motor action (left or right), previous trial reward outcome (+10, 0, or -10), and previous trial RPE as independent variables (regressors), neuron ID as the random intercept, and average FR for each neuron during the expectation period as dependent variable:

$$\mathrm{FR} \sim 1 + \text{previous motor\_action} + \text{previous\_reward outcome} + \text{previous\_trial\_RPE} + (1|\text{neuron})$$

To examine if the reward outcomes should be split by block type, we carried out a separate GLM that included an interaction term between previous reward outcome and block type on expectation period firing rate:

$$\mathrm{FR} \sim 1 + \text{previous\_reward outcome} * \text{condition}$$

We also used GLM analysis to examine the effect of the independent variables motor action (left or right), reward outcome (+10, 0, or -10) and RPE on the dependent variable of the feedback firing rate, specified as:

$$\mathrm{FR}\,(\text{feedback period}) \sim 1 + \text{motor\_action} + \text{reward outcome} + \text{RPE}$$

Finally, we also used a mixed effect GLM model to examine the relationship between FRs and RTs, including block type as fixed effect and patient ID as random effects:

$$\mathrm{RT} \sim 1 + \text{previous FR} + \text{block type} + (1 \mid \text{patient ID})$$

## QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details can be found in the results section as well as the respective figure legends. Data was reported as mean ± standard deviation. Various statistical tests were used based on the best fit for the data. Averages data for patients or neurons across different conditions were analyzed using paired *t* test (Figures 2, 3, and 4A). Individual trial data was analyzed using a linear model to examine significance across multiple variables. A non-parametric Wilcoxon rank-sum test was used for not evenly distributed data in the reaction time analysis (Figure 4C). For examining significant time periods, permutation cluster testing was used. Significance across all statistical tests was determined by $p < 0.05$. All statistical analysis was done in MATLAB.