NEUROSCIENCE

Subsecond fluctuations in extracellular dopamine encode reward and punishment prediction errors in humans

L. Paul Sands[1,2]†‡, Angela Jiang[2]‡, Brittany Liebenow[1,2]‡, Emily DiMarco[1,2], Adrian W. Laxton[3], Stephen B. Tatter[3], P. Read Montague[4,5,6], Kenneth T. Kishida[1,2,3]*

In the mammalian brain, midbrain dopamine neuron activity is hypothesized to encode reward prediction errors that promote learning and guide behavior by causing rapid changes in dopamine levels in target brain regions. This hypothesis (and alternatives regarding dopamine's role in punishment-learning) has limited direct evidence in humans. We report intracranial, subsecond measurements of dopamine release in human striatum measured, while volunteers (i.e., patients undergoing deep brain stimulation surgery) performed a probabilistic reward and punishment learning choice task designed to test whether dopamine release encodes only reward prediction errors or whether dopamine release may also encode adaptive punishment learning signals. Results demonstrate that extracellular dopamine levels can encode both reward and punishment prediction errors within distinct time intervals via independent valence-specific pathways in the human brain.

INTRODUCTION

Dopamine neurons are critical for mammalian brain function and behavior (*1*), with changes in dopaminergic efficacy believed to underlie a wide range of human brain disorders including substance use disorders, depression, and Parkinson's disease (*2*–*5*). Midbrain dopamine neurons project to the basal ganglia, cortical brain regions associated with cognitive, limbic, and motor functions, and form recurrent connections with ventral and dorsal striatum (*6*). In this way, dopamine signaling is thought to influence different distributed brain networks that control processes supporting stimulus-driven and goal-directed decision-making, such as reward learning, motor planning and execution, motivation, and emotion (*1*, *6*). A leading hypothesis that connects these disparate computational roles for dopamine proposes that dopamine neurons encode information about errors in an organism's expectations about rewarding outcomes, so-called reward prediction errors [RPE; (*7*, *8*)]. Specifically, in nonhuman animal research, it has been shown that phasic changes in dopamine neuron spiking activity encode "temporal difference" RPEs [TD-RPEs; (*7*–*14*)], an optimal learning signal derived within computational reinforcement learning theory (*15*) and that has recently been central to major advances in the development of deep learning artificial neural networks capable of autonomously achieving human expert-level performance on a variety of tasks (*16*–*19*).

Decades of nonhuman animal research support the idea that dopamine neurons signal RPEs in the mammalian brain [(*7*–*14*); see (*11*) for review]; however, in humans, direct evidence is limited. There is clear evidence in humans that changes in the firing rate of putative dopamine neurons encode RPEs (*20*), and regions rich in afferent dopaminergic input show changes in blood oxygen–level–dependent signals consistent with physiological processing of RPEs (*21*–*23*). Still, using indirect measures of dopamine release like spiking rates at the cell body or hemodynamic signals does not provide direct evidence that dopamine release in target regions effectively signals RPEs. In rodents, subsecond changes in extracellular dopamine levels in the striatum have been measured using fast-scan cyclic voltammetry (FSCV) and rapid-acting, genetically encoded, fluorescent dopamine sensors [e.g., dLight and GRAB; (*24*, *25*)]. These studies reveal that dopamine levels not only reflect RPEs (*12*–*14*) but also respond to diverse affective stimuli [e.g., drug-predictive cues; (*26*, *27*)] and vary with specific recording location (*28*) and task demands [e.g., effort costs; (*29*)]. Consistent with this, rodent and nonhuman primate studies have shown that changes in dopamine neuron firing rate may also encode aversive prediction errors (*13*, *30*–*35*). Relatedly, human functional magnetic resonance imaging experiments suggest that RPE and punishment prediction error (PPE) signals are represented in dopamine-rich regions during learning about appetitive and aversive outcomes (*36*–*39*).

Recently, studies leveraging the ability to directly measure dopamine release in the human brain with high temporal resolution have revealed that subsecond changes in dopamine levels reflect both actual and counterfactual error signals during risky decision-making (*40*, *41*), the average value of reward following a sequence of decisions (*42*), and nonreinforced, although goal-directed, perceptual decision-making (*43*). In experiments where RPEs could be estimated (*40*, *41*), dopamine levels seemed to entangle actual and counterfactual information (i.e., outcomes that "could have been" had a different choice been made) for both gains and losses, resulting in a superposed value prediction error signal (*40*). These results suggest the hypothesis that reward and punishment information, encoded by extracellular dopamine

[1]Neuroscience Graduate Program, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA. [2]Department of Physiology and Pharmacology, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA. [3]Department of Neurosurgery, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA. [4]Wellcome Centre for Human Neuroimaging, University College London, WC1N 3BG London, UK. [5]Fralin Biomedical Research Institute, Virginia Tech, Roanoke, VA 24016, USA. [6]Department of Physics, Virginia Tech, Blacksburg, VA 24061, USA.
*Corresponding author. Email: kkishida@wakehealth.edu
†Fralin Biomedical Research Institute at Virginia Tech, Roanoke, VA 24016, USA.
‡These authors contributed equally to this work.

fluctuations, could be derived from independent streams but combined or differentiated by downstream neurons in the striatum (*44*).

We sought to determine whether dopamine release in human striatum specifically encodes TD-RPEs in humans as initially suggested by work in nonhuman primates (*7*, *8*). We also sought to test an alternative hypothesis that dopamine release in these same loci also encodes PPEs, the possibility of which remains debated (*13*, *30*–*35*). To test these hypotheses, we used human voltametric methods (Fig. 1A) (*40*–*43*), while participants performed a decision-making task (Fig. 1B) that allowed us to disentangle the impact of rewarding and punishing feedback on dopamine release and choice behavior. This approach allowed us to monitor rapid phasic changes in dopamine levels (Fig. 1C), while participants learned from rewarding as well as punishing feedback. The specific task design allowed us to test two different reinforcement learning models that express the mutually exclusive hypotheses that dopamine release encodes RPEs and PPEs via (i) a unidimensional valence system versus (ii) a valence-partitioned system (*45*, *46*), whereby appetitive and aversive stimuli are processed by independent systems, thereby allowing learning of co-occurring although statistically independent appetitive and aversive stimuli (fig. S1).

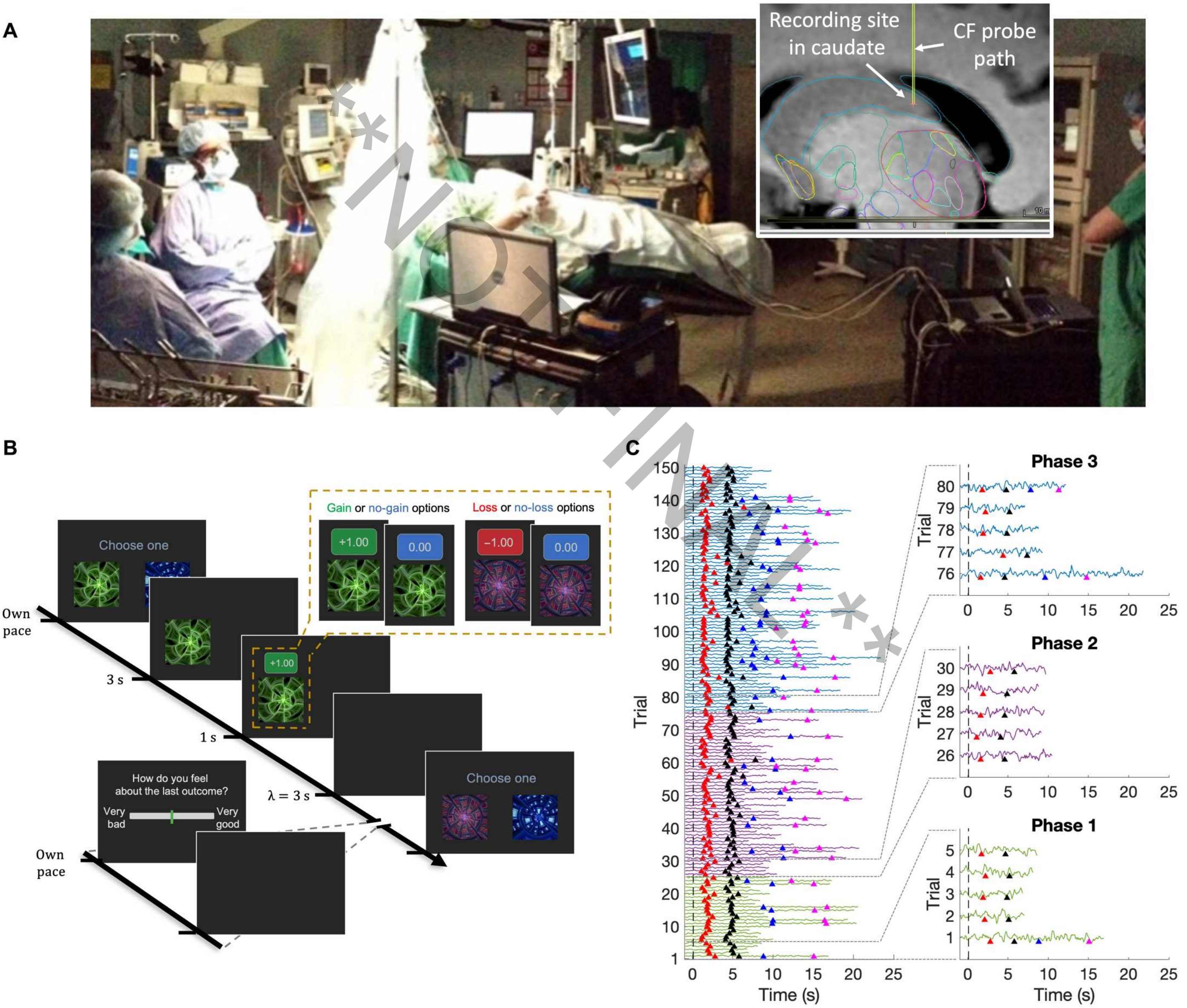

Fig. 1. Probabilistic reward and punishment task and associated trial-by-trial dopamine time series recorded via human voltammetry. (**A**) Blurred (to protect privacy) photograph depicting the operating room during a human voltammetry experiment during DBS neurosurgery; inset: MRI showing general location of the research-use carbon-fiber (CF) microelectrode within the caudate. (**B**) Schematic of a trial from the choice task. (**C**) Trial-by-trial time series of caudate dopamine levels recorded from a single participant, with time series colored according to task phase; vertical dashed line indicates when the choice options were presented on each trial, and colored markers indicate trial events of interest (red, choice onset; black, outcome onset; blue, rating query onset; magenta, rating submitted).

RESULTS

Human voltammetry experimental design

Participants (n = 3) were adult patients diagnosed with essential tremor (ET) who consented to undergo deep brain stimulation (DBS) electrode implantation neurosurgery (Fig. 1A). Before the day of surgery, all participants provided written informed consent to participate in the research procedure after deciding to undergo the clinical procedure. The neuroanatomical target of DBS lead implantation surgery for patients with ET is the ventralis intermediate nucleus (VIM) of the thalamus; this surgery permits carbon-fiber microelectrode recording within the caudate nucleus, a major site for dopaminergic innervation and dopamine release (Fig. 1A, inset). Notably, the pathophysiology of ET is thought to not involve disruptions of the dopaminergic system (*47*). Before implanting the DBS lead, a carbon-fiber microelectrode is used for voltammetric recordings along the trajectory that the DBS lead may be placed (*40*–*43*). In the present work, the carbon-fiber microelectrode was placed in the caudate (generally in the medial-posterior regions; see inset in Fig. 1A), and dopamine measurements were sampled once every 100 ms, while participants performed the reward and punishment learning task. Following the research procedure, the carbon-fiber microelectrode is removed, and the DBS electrode implantation surgery is completed. No change in the outcome or associated risks have been associated with performing this kind of intracranial research (*48*).

The behavioral task that we used is a probabilistic reward and punishment (PRP) learning task with reversal learning where participants' actions were reinforced or punished with monetary gains or losses. Participants are instructed and actually paid a bonus according to the dollar amounts that they earn in the task. Unbeknownst to the participants, the task is setup in stages (fig. S2), such that the initial stage (phase 1) is biased toward probabilistic gain trials (binary outcomes, \$1 or \$0) where participants can earn an initial reserve of cash before entering phase 2, which introduces trials with probabilistic losses (binary outcomes, −\$1 or \$0). In the final stage (phase 3), the probabilities of gain or loss outcomes associated with the choice cues are held constant, but the magnitudes of the outcomes are changed, such that the expected values change which options should be expected to pay the most or least (fig. S2). Optimal performance on this task requires participants to learn from positive and negative feedback to select the option on each trial that maximizes the expected reward and minimizes the expected punishment.

Human dopamine levels and TD-RPEs

Behavioral data demonstrated that patients with ET learned the PRP task's incentive structure: Overall, they chose the best option on a given trial more often than chance. Moreover, patients with ET did not perform significantly differently from a control cohort of human adults performing the same task in a behavioral laboratory setting in a manner that would suggest any specific additional difficulty with the experimental task for patients with ET (fig. S3). To test whether subsecond dopamine fluctuations in human caudate reflected TD-RPEs, we extracted time series of dopamine levels

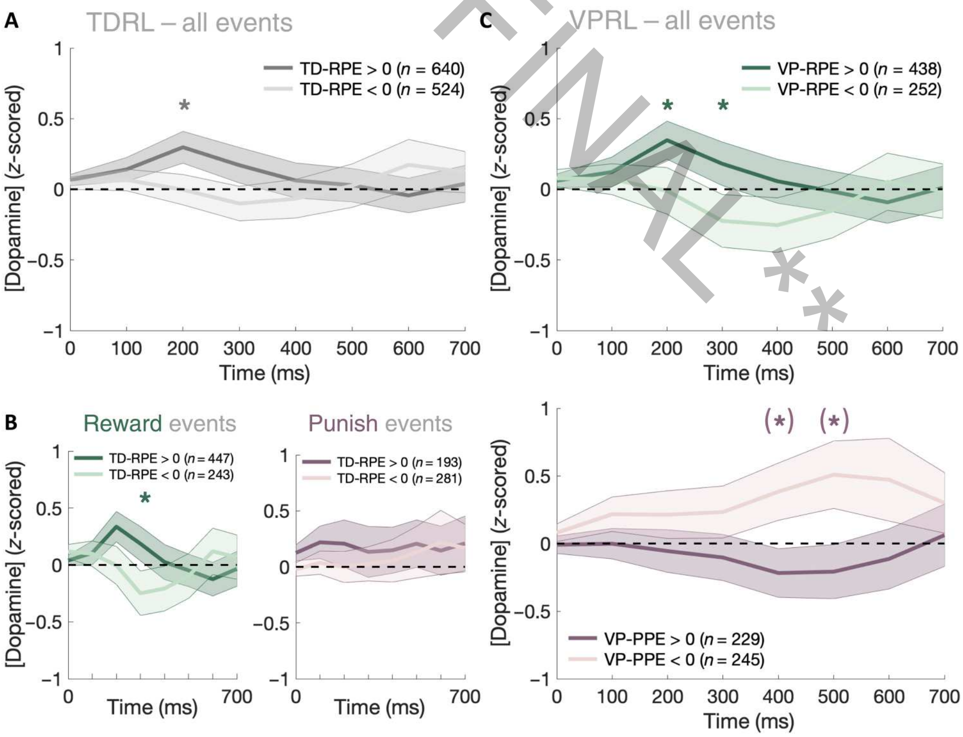

Fig. 2. Phasic dopamine levels in human caudate reflect valence-partitioned RPEs and PPEs. Dopamine responses from 0 to 700 ms following prediction errors events in the PRP task are categorized by trial type and prediction error sign. (**A**) Statistically, phasic dopamine transients fail to separate positive and negative TD-RPEs. (**B**) Dopaminergic TD-RPE responses sorted by trial type: gain-expected trials (left) and loss-expected trials (right). (**C**) Phasic dopamine transients across all trials sorted by VP-RPE sign (top) and VP-PPE sign (bottom). Asterisks denote permutation test, $P < 0.05$; parentheses around asterisks denote significance after correcting permutation tests for multiple comparisons following the Benjamini-Hochberg procedure.

on each trial aligned to the moments of option presentation, action selection, and outcome presentation, each of which were expected to elicit TD-RPEs during the course of the task. With learning, TD prediction error signals gradually shift in time from occurring in response to previously unexpected outcomes to instead occurring in response to the presentation of outcome-predictive stimuli (*7*, *8*, *14*, *15*). This is a hallmark characteristic of TD learning processes (*14*) that explains higher-order conditioning behaviors and temporal patterns of dopamine neuron activity that other theoretical models that do not consider time within a trial [e.g., Rescorla-Wagner rule; (*49*)] do not account for (*6*, *7*, *14*). Thus, we hypothesized that dopamine fluctuations would signal TD-RPEs in response to events within single trials related to the moment of option presentation, action selection, and outcome presentation.

We fit a temporal difference reinforcement learning (TDRL) model to participant behavior and compared the average dopamine timeseries estimates for positive TD-RPEs (n = 640) and negative TD-RPEs (n = 524) (Fig. 2, A and B, figs. S4 and S5). Prior studies in humans, rodents, and nonhuman primates demonstrate that changes in midbrain dopamine neuron firing rate in response to rewards and punishments and their predictive cues occur within 100 to 700 ms after event and have nonlinear, multiphasic temporal patterns that differ between rewards and punishments (*7*, *8*, *10*, *20*, *30*–*35*). Moreover, phasic changes in extracellular dopamine levels in human striatum that reflect gain and loss prediction error signals during risky decision-making have been observed within 200 to 600 ms after event (*40*, *41*). Accordingly, we hypothesized that dopamine-based RPE and PPE signals ought to occur within a similar time window and focused our analyses on dopamine responses from 0 to 700 ms following option presentation, action selection, and outcome presentation.

We found that, across all events from all trials, subsecond dopamine fluctuations in human caudate did not significantly distinguish positive versus negative TD-RPEs [two-way analysis of variance (ANOVA) (RPE sign, 0 to 700 ms): $F_{\text{RPE-sign}}(1,7) = 1.11$, $P = 0.29$; Fig. 2A]. However, the dopamine TD-RPE response was not linear over time; splitting the analysis window into early (0 to 300 ms) and late (400 to 700 ms) phasic components of the event-related dopamine responses (*34*) showed that dopamine fluctuations across all trials did distinguish TD-RPEs at the early component but not the late components [early component two-way ANOVA (RPE sign, 0 to 300 ms: $F_{\text{RPE-sign}}(1,3) = 6.11$, $P = 0.013$; late component two-way ANOVA (RPE sign, 400 to 700 ms): $F_{\text{RPE-sign}}(1,3) = 0.18$, $P = 0.67$]. Post hoc two-sample t tests and nonparametric permutation tests run for each individual time sample 0 to 700 ms after RPE indicated a significant difference between positive and negative TD-RPEs across all trials at 200 ms {right-tailed two-sample t test [(RPE > 0) > (RPE < 0)]: $t_{200\text{ms}}(1162) = 1.90$, P = 0.029; permutation test: $P = 0.026$; fig. S4}. We next further separated dopamine responses into reward and punishment trial–specific prediction errors (Fig. 2B), which revealed that dopamine release weakly distinguished TD-RPEs on reward trials within the first 300 ms after event {early component two-way ANOVA: $F_{\text{RPE-sign}}(1,3) = 3.46$, $P = 0.063$; permutation test: $P_{300\text{ms}} = 0.036$; right-tailed two-sample t tests [(RPE > 0) > (RPE < 0)]: $t_{300\text{ms}}(688)$ = 1.74, $P = 0.041$; Fig. 2B} but did not distinguish TD-RPEs on punishment trials [early component two-way ANOVA: $F_{\text{RPE-sign}}(1,3)$ = 0.04, $P = 0.84$; late component two-way ANOVA: $F_{\text{RPE-sign}}(1,3)$ = 2.51, $P = 0.11$; Fig. 2B]. Collectively, these results indicate that phasic fluctuations in dopamine levels in human caudate may signal TD-RPEs in the first 200 to 300 ms after event, although this is only the case for reward-related actions and expected outcomes.

Human dopamine levels and valence-partitioned prediction errors

Prior work demonstrated that dopaminergic responses could track PPEs (*13*, *31*–*35*), but results shown in Fig.2B suggest that dopamine fluctuations do not reflect temporal difference reward learning when the outcome stimulus is punishing or expected to be punishing (e.g., monetary losses). Thus, we hypothesized that dopamine may encode PPEs but as an independent, punishment-specific valuation system (*44*–*46*). We tested this hypothesis by fitting to participant behavior a "covalent learning" model based on a valence-partitioned reinforcement learning (VPRL) framework that expresses the independence of reward and punishment learning explicitly (fig. S1) (*44*–*46*). This is accomplished by hypothesizing that two separate neural systems implement TD learning that individually process either rewarding (positive valence system) or punishing (negative valence system) information. We note that the reward system (i.e., "positive valence" system) in this VPRL framework serves the same role as dopamine-based TD-RPE signaling, except that the positive system treats negative outcomes as though nothing happened (i.e., actual losses are treated by the positive system as an outcome equal to zero); the crux of the covalent learning model is the hypothesized existence of a separate, parallel neural system dedicated to learning the statistical structure of negatively valent information using a TD learning algorithm, which would allow optimal learning of statistically independent rewarding and punishing events. Whereas dynamics within each valence system independently govern either reward or punishment learning, signals from both systems can be integrated by downstream systems [e.g., brain regions to furnish more complex computations underlying affective behaviors (fig. S1), such as signaling the net affective valence of a stimulus or action (i.e., contrast: positive system signals minus negative system signals) or its degree of behavioral excitation (i.e., arousal via integration: positive system signals plus negative system signals)].

Fitting subjects' behavior to a VPRL model resulted in a better fit to participant behavior compared to TDRL (table S1). We found that these results replicated in an independent cohort of healthy human adults (N = 42) who completed the PRP task on a computer in a behavioral laboratory setting (table S1 and figs. S3, S6, and S7) (*45*). Further comparisons revealed that VPRL algorithms may perform reward and punishment learning more efficiently (e.g., learns about punishment structure faster) than traditional TDRL models that do not partition appetitive and aversive stimuli (figs. S6 and S7). We next tested the hypothesis that dopamine release encoded valence-partitioned RPEs (VP-RPEs) and valence-partitioned PPEs (VP-PPEs) by sorting dopamine release time series data by the VPRL model–specified prediction errors: positive VP-RPEs (n = 438), negative VP-RPEs (n = 252), positive VP-PPEs (n = 229), or negative VP-PPEs (n = 245) (Fig. 2C and figs. S4 and S5). We found that dopamine transients distinguished VP-RPEs on reward trials within the same time window as found for TD-RPEs {early component two-way ANOVA: $F_{\text{RPE-sign}}(1,3) = 4.51$, $P = 0.034$; permutation test: $P_{200\text{ms}} = 0.047$, $P_{300\text{ms}} = 0.046$; right-tailed two-sample t test [(RPE > 0) > (RPE < 0)]: $t_{200\text{ms}}(688) = 1.65$, P =

0.049; $t_{300\text{ms}}(688) = 1.63$, $P = 0.051$; Fig. 2C}. However, notably, we also observed that phasic dopamine responses effectively distinguished VP-PPE signals within a temporal window distinct from VP-RPE responses, lasting from 400 to 500 ms following a prediction error {late component two-way ANOVA: $F_{\text{RPE-sign}}(1,3) = 10.7$, $P = 0.001$; permutation test: $P_{400\text{ms}} = 0.014$, $P_{500\text{ms}} = 9.8 \times 10^{-3}$; left-tailed independent sample t test [(PPE > 0) < (PPE < 0)]: $t_{400\text{ms}}(472) = -2.3$, $P = 0.010$; $t_{500\text{ms}}(472) = -1.80$, $P = 0.036$; Fig. 2C}. These results demonstrate that subsecond dopamine fluctuations in human caudate may encode VP-RPEs and VP-PPEs.

Last, we tested a theoretical prediction of the VPRL framework (*44*–*47*) that an action or stimulus's overall affective valence (i.e., reinforcement or punishment) and behavioral activation (i.e., high/low arousal) could be represented by the difference and sum of VP-RPE and VP-PPE signals, respectively (fig. S1). In traditional TDRL, it is often assumed that affective valence and behavioral activation are both simply defined by whether a TD-RPE is positive or negative: Positive TD-RPEs are reinforcing and tend to increase behavioral activation, and negative TD-RPEs are punishing and tend to inhibit behavioral activation; our results demonstrate that this information is only distinguishable by the TDRL hypothesis on reward trials (Fig. 2B). In comparison, it has been suggested that a dual-valence system framework like VPRL could provide greater resolution for both valence contrast and behavioral activation information processing (*44*–*47*). Thus, we tested whether integrated VPRL error signals could encode these signals and found significant differences between dopamine responses representing positive affective valence (i.e., reinforcement; average across VP-RPEs > 0 and VP-PPEs < 0) and negative affective valence (i.e., punishment; average across VP-RPEs < 0 and VP-PPEs > 0), specifically between 200 and 500 ms after event (fig. S8, A and C), although there were no significant differences in dopamine responses representing behavioral activation signals (fig. S8, B and D).

Decoding RPEs and PPEs

Fluctuations in extracellular dopamine levels are expected to provide a decodable signal to downstream neural structures. To determine whether the signals that we report (Fig. 2C) are robust enough to be decoded, we trained logistic classifiers to distinguish dopamine time series resulting from positive and negative prediction errors on reward trials (Fig. 3, A and B) or positive and negative prediction errors on punishment trials (Fig. 3, C and D). The classifiers trained to discriminate positive versus negative RPEs (TD-RPEs or VP-RPEs on rewarded trials) performed comparably for both TDRL and VPRL models (Fig. 3, A and B). Conversely, classifiers trained to discriminate positive from negative PPEs (TD-RPEs or VP-PPEs on punishment trials) only succeeded when the dopamine time series were parsed according to the VPRL model and performed at chance level when the dopamine transients were hypothesized to be encoded by TDRL (Fig. 3, C and D).

DISCUSSION

We demonstrate, in humans, that subsecond dopamine fluctuations in the caudate nucleus reflect RPE and PPE signals as predicted by a VPRL framework. Collectively, our results suggest that human decision-making is influenced by independent, parallel processing of appetitive and aversive experiences and expectations that can affect modulation of dopamine release in striatal regions on rapid time scales (hundreds of milliseconds). Our findings provide an account for previous observations that dopamine fluctuations in human striatum appear to superpose actual and counterfactual information related to gains and losses during risky decision-making (*40*, *41*) and point toward a candidate neurochemical substrate for observed hemodynamic changes in human striatum related to appetitive and aversive learning processes (*21*–*23*, *36*–*39*).

Related ideas have been proposed, for instance, that rewards and punishments are integrated together during learning (as opposed to being processed independently), leading to a "zero-sum" prediction error that is signaled by dopamine neurons only if the prediction error is positive [i.e., rewarding; (*50*)] or that positive and negative RPEs are learned about "asymmetrically" [i.e., different learning rates; (*51*, *52*)]. Notably, however, these models propose that the brain learns a single action value representation updated via a single prediction error signal that integrates over rewards and punishments. We emphasize that such proposals are distinct from what is proposed by the VPRL framework, where reward and punishment learning are performed simultaneously, in parallel, by independent neural systems that generate two valence-specific TD prediction error signals for updating separate reward- and punishment-specific action value representations (*45*, *46*). We note that VPRL is compatible with recently proposed distributional reinforcement learning methods (*52*), but, to our knowledge, explicit representations of distinct distributions for punishment learning (valence partitioning) have not yet been explored in in these distributional approaches.

Averaging dopamine time series across all patients revealed that dopamine transients reflect VP-RPEs and VP-PPEs, distinguishing between positive and negative VP-RPEs and VP-PPEs within distinct temporal windows. Instead of averaging dopamine responses for all events across all patients, if we first average dopamine responses across all trials within each patient and then average across participants at the group-level (i.e., mixed-effects analysis), then we obtain similar results (figs. S5 and S8). However, we note that the present human voltammetry data may violate assumptions underlying random- and mixed-effects analyses, in that there remains a significant gap in knowledge about the sources of variability in the measured signals in these relatively early experiments. For example, unlike nonhuman primate and rodent studies, our human volunteers come to the clinic with significant variation in life experience and genetic background. Furthermore, the data collected are expected to vary with the microanatomical environment of the recording electrode (the active surface is roughly a cylinder of only 110 μm in length by 7 μm in diameter) within the caudate (a volume of approximately 4000 mm^3) of each patient. Different subregions of the primate caudate (e.g., head, body, and tail subdivisions; matrix and patch subcompartments) are topographically organized to perform distinct functions as components of specific but interacting whole-brain networks (*6*). Moreso, local regulation of dopamine release along axonal projections, independent of signals generated at dopaminergic soma, is known to cause variations in subsecond dopamine fluctuations at release sites (*53*). How these and other sources of variation influence the signal that we record is not at present known, but studies like this one and others (*40*–*43*) are providing our first glimpses at what kinds of signals can be reflected in subsecond dopamine fluctuations in humans. Accordingly, a fixed-effects analysis (Fig. 2) is more sensitive for describing the structure in observed dopamine responses

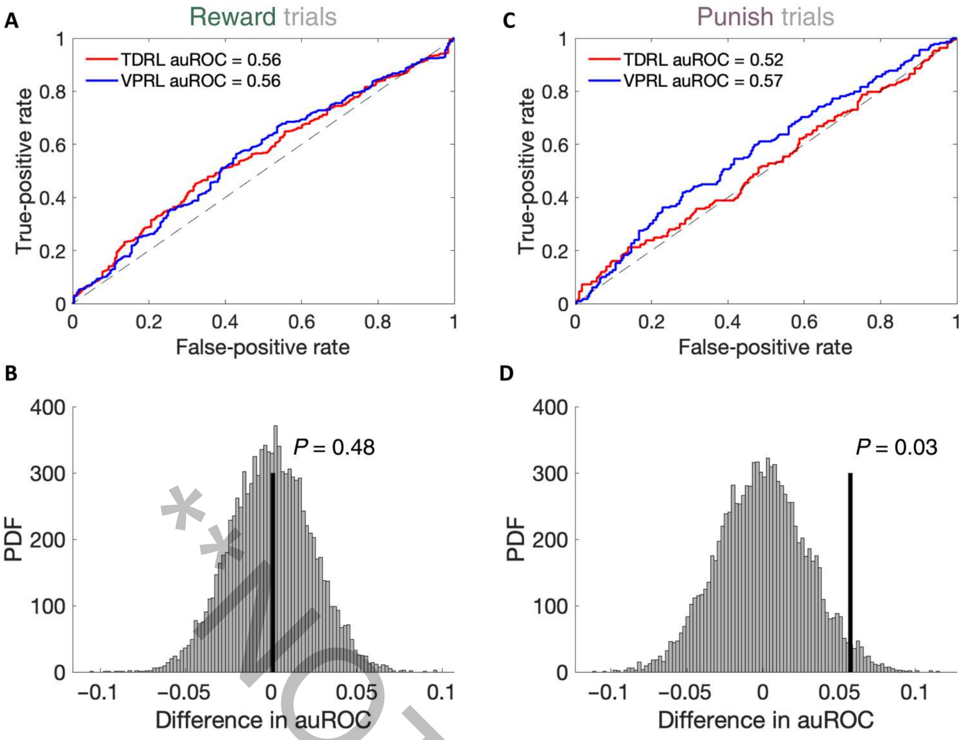

Fig. 3. VPRL RPEs and PPEs can be decoded from human dopamine transients. Performance of the logistic classifier trained on (**A**) TDRL-derived (red) or VPRL-derived (blue) positive and negative RPEs is comparable across models, with (**B**) the difference in the area under the receiver operating characteristic (auROC) curve values not being statistically significant; *P* value derived from permutation test with 10,000 iterations. (**C** and **D**) Same as (A) and (B) but for punishment trials; the difference in auROC values for the PPE logistic classifiers was significantly different for TDRL and VPRL ($P = 0.03$). PDF, Posterior Density Function: Count.

common across all observations reported (i.e., implicitly averaged over caudate subregions), although less committed to the nature of signals expected to be observed in future work.

Our observation that early (0 to 300 ms) dopamine responses signaled RPEs and later (400 to 700 ms) responses signaled PPEs is generally consistent with the timing of activity within neuroanatomical circuitry by which rewarding and aversive stimuli modulates dopamine neuron activity to effect associative learning (*7*, *8*, *10*, *20*, *30*–*35*, *54*–*56*). For instance, brain regions such as dorsolateral and central gray pontine tegmental nuclei and their circuit interactions with the habenula and hypothalamus could feasibly communicate valence-specific information to midbrain dopamine neurons to affect prediction error responses at time scales relevant to our present findings (*57*–*59*). More generally, the apparent temporal coding of valence information by dopamine levels in the caudate (Fig. 2C and fig. S8) suggests that how basal ganglia and distributed cortical brain regions differentially regulate goal-directed decision-making depends on the temporal covariance of dopamine (and other neuromodulators') fluctuations. Still, further research is necessary to probe the possible cellular or circuit mechanisms giving rise to the observed outcome valence-dependent temporal patterns in dopamine prediction error responses. For example, combining recordings of dopaminergic action potentials at the cell body and neurotransmitter release at target brain regions could test more comprehensively whether distinct subpopulations of dopamine neurons may be activated to signal valence-specific prediction errors or whether a separate neural system controls, more locally, the timing and polarity of dopaminergic activity in response to valent behavioral reinforcers.

The approach used to collect the data presented here are constrained by the requirement of standard-of-care neurosurgical procedures that provide access deep into the human brain. Nonetheless, our data suggest an extension of traditional TDRL-based algorithms for understanding how humans process affective information and the role dopamine signals may play in the human brain. Furthermore, these kinds of data are invaluable for investigating and understanding how neuromodulators like dopamine affect human behavior, human decision-making, and human subjective experience. Studies and results like those presented here may influence how we understand and think about the role dopamine that plays in psychiatric and neurological disorders, where, in many instances, the extant body of knowledge about dopamine in humans is restricted to measurements that occur on temporal resolution of the order of several minutes to days. Future work may allow further direct investigation of how subsecond dopamine signals are altered in humans with ET and comorbid psychiatric conditions or how rapid dopaminergic signals may be altered in patients with Parkinson's disease with or without comorbid psychiatric and nonmotor Parkinson's disease–related symptoms.

While much work in translational and basic neuroscience is focused on basic biological mechanisms in model organisms, it remains unclear whether such nonhuman models are appropriate systems to investigate fundamentally human phenomena like subjective affect, human-level willful choice, and related psychiatric and neurological disorders. The present study confirms prior discoveries made in nonhuman model organism pertaining to dopamine's role in encoding RPEs (*7*–*14*) but also shows that this hypothesis, previously supported by decades of research in model organisms, is incomplete regarding dopamine's role in encoding aversive feedback. We show that dopamine does not just respond to aversive stimuli but does so in a manner analogous to TD learning but specific aversive outcomes. We note that it has not escaped our notice that the PPE aversive learning system has features that suggest a role in anxiety disorders analogous to models of the RPE reinforcement learning system role in addiction, depression, and obsessive-compulsive disorder.

MATERIALS AND METHODS

Patient recruitment and informed consent

A total of 11 patients (6 females and 5 males; age range, 48 to 82; mean age ± SD, 67.5 ± 10.9) diagnosed with ET and approved candidates for DBS treatment participated in this study. Of the 11 patients, a subset of three patients performed the procedure while carbon-fiber microelectrodes recorded dopamine release in their caudate, and one patient performed the procedure while a carbon-fiber microelectrode recorded dopamine release in their thalamic VIM. The other seven patients performed the task while recordings were made with a tungsten microelectrode. While all 11 patients' behavioral data were included in analyses for hierarchical parameter estimation, the tungsten microelectrode ($n = 7$) and thalamic VIM ($n = 1$) neurochemical recordings were not included in the present work. Thus, all 11 patients' behavioral data were used for the computational modeling and behavioral analyses (figs. S3, S6, S7, and S9), whereas only three patients' neurochemical recordings (with carbon-fiber electrodes in caudate) were included in the dopamine prediction error analyses (Figs. 2 and 3 and figs. S4, S5, and S8).

After informed written consent was obtained from each patient, patients were given details about the decision-making task (i.e., PRP task) and were familiarized with the type of outcomes experienced during game play and the controllers used for submitting responses. The experiment was approved by the Institutional Review Board (no. IRB00017138) of Wake Forest University Health Sciences (WFUHS). Of the 11 patients that participated in the study, four patients did not complete all 150 trials of the task (range, 121 to 148 trials).

In addition to the cohort of patients with ET, a behavior-only cohort of healthy adult humans ($N = 42$; 19 females) was recruited from the local Winston-Salem community to complete the PRP task. Informed written consent was obtained from each participant, and the experiment was approved by the IRB (no. IRB00042265) of WFUHS. All behavioral experiments were conducted at WFUHS.

PRP task experimental procedure

The PRP task (Fig. 1B and fig. S2) is a 150-trial, two-choice monetary reward and punishment learning task, where chosen options are reinforced probabilistically with either monetary gains (or no gain) or monetary losses (or no loss). Six options (represented by fractal images) comprise the set of possible actions, with each option assigned to one of the three outcome probabilities (25, 50, and 75%) and one of the two outcome valences (monetary gain or loss); thus, there are three reward-associated "gain/no-gain" options and three "loss/no-loss" options in the task, and the assignment of options to outcome probabilities and valences is randomized across participants. On each trial, two of the six options are presented (note that option pairings are random, not fixed); depending on the phase of the task (phase 1, trials 1 to 25; phase 2, trials 26 to 75; and phase 3, trials 76 to 150), either two of the three gain/no-gain options are presented (i.e., gain/no-gain trials), two of the three loss/no-loss options are presented (i.e., loss/no-loss trials), or one of each gain/no-gain and loss/no-loss options are presented (i.e., "mixed" trials). Participants were told that certain options in the PRP task would earn them money and some options would lose them money, and participants were instructed that their goal was to maximize their earnings on the task and that they would receive their total earnings as a bonus monetary payment at the end of the study visit.

At the beginning of the experiment (phase 1, trials 1 to 25), each trial starts with the presentation of two of the three possible gain/no-gain options, and participants are reinforced with either a monetary gain or nothing ($1 or $0) according to the chosen option's fixed probability. In phase 2 (trials 26 to 75), the task introduces loss/no-loss trials that present two of the three loss/no-loss options that result in either a monetary loss or nothing (−$1 or $0) with fixed probabilities. In this phase, there are an equal number of gain/no-gain and loss/no-loss trials, randomly ordered. In phase 3 (trials 76 to 150), two options are presented randomly such that any trial may consist of two gain/no-gain options and two loss/no-loss options or one gain/no-gain and one loss/no-loss option. Moreover, in phase 3, the outcome magnitudes of all options change such that the 25, 50, and 75% "gain" options now payout $2.50, $1.50, and $0.50, respectively, and the 25, 50, and 75% "loss" options now lose −$1.25, −$0.75, and −$0.25, respectively (see dashed lines in fig. S2).

On each trial, participants select an option at their own pace. Once a selection has been made, the unchosen option disappears at the same time that the chosen option is highlighted, and this screen lasts for 3 s. The outcome is then displayed for 1 s followed by a blank screen that lasts for a random time interval (defined by a Poison distribution with $\lambda = 3$ s) before the next trial begins. In addition, on each trial with probability of 0.33, the blank screen following the outcome presentation is followed by a subjective feeling rating screen that consists of the text "How do you feel about the last outcome?" and a visual-analog rating scale with a vertical bar cursor that can be moved by the participant. Participants are asked to rate their feelings about the experienced outcome with this visual-digital scale, after which the blank screen reappears for another random time interval before the next trial begins.

Behavioral data analysis

TDRL model

In the standard TDRL model (*15*, *60*), the expected value of a state-action pair $Q(s_i, a_i)$, where i indexes discrete time points in a trial, is updated following selection of action a_i in state s_i according to

$$Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \alpha\delta_i \tag{1}$$

where $0 < \alpha < 1$ is a learning rate parameter that determines the weight prediction errors have on updating expected values and δ_i is the TD-RPE term

$$\delta_i = [\text{outcome}_i + \gamma \max_{a} Q(s_{i+1}, \tilde{a})] - Q(s_i, a_i) \tag{2}$$

where outcome$_i$ is the outcome (positive or negative) experienced in state s_i after taking action a_i, $0 < \gamma < 1$ is a temporal discount parameter that discounts outcomes expected in the future relative to immediate outcomes, and $\max_a Q(s_{i+1}, \tilde{a})$ is the maximum expected action value over all actions $\tilde{a}$ afforded in the next state s_{i+1}. We defined the trials of the PRP task as consisting of $i = \{1,2,3,4\}$ event time points [1, options presented; 2, action taken; 3, outcome presented; and 4, (terminal) transition screen]. We modeled participant choices (choice$_t$) on each trial t of the PRP task with a softmax choice policy (i.e., categorical logit choice model) that assigns probability to choosing each of the two options presented on a trial according to the learned Q values of the two options. For example, for a trial that presents option 2 and option 5, the corresponding action values at the moment of option presentation, $Q(s_1, \text{opt\_2})$ and $Q(s_1, \text{opt\_5})$, are used to compute the probability of selecting each option

$$P[\text{choice}_t = \text{opt\_2} \mid Q(s_1, \text{opt\_2}), Q(s_1, \text{opt\_5})] = \frac{e^{Q(s_1,\text{opt\_2})/\tau}}{e^{Q(s_1,\text{opt\_2})/\tau} + e^{Q(s_1,\text{opt\_5})/\tau}} \tag{3}$$

where $0 < \tau < 20$ is a choice temperature parameter that determines the softmax function slope and parameterizes an exploration versus exploitation trade-off where higher temperature values lead to a more randomized choice selection policy and lower temperature values lead to a more winner-take-all, deterministic choice policy.

VPRL model

For VPRL (*45*, *46*), we extend the standard TDRL framework by specifying that two separate value representations are learned for each action, corresponding to the rewarding value and punishing value of each action, and that separate (neural) systems signal reward- and punishment-specific prediction errors to update the reward- and punishment-associated action values, respectively. In this way, VPRL treats "positive" (*P*) and "negative" (*N*) outcomes as though separate, parallel *P*- and *N*-systems effectively establish a partition between the processing of rewarding and punishing outcomes. *P*- and *N*-system action values are estimated (Q^P and Q^N, respectively) independently, although each system learns these outcome valence-specific action values using temporal difference learning (see Eqs. 4 to 7). We model the integration of Q^P and Q^N in the simplest manner (i.e., subtraction; Eq. 8) when value-based decisions must be made, although alternative approaches for integrating these value estimates may be investigated in future work.

In VPRL, *P*- and *N*-systems update action value representations via TD prediction errors on every episode but by valence-specific rules (*P*-system, Eq. 4; and *N*-system, Eq. 5). The *P*-system only tracks rewarding (i.e., appetitive) outcomes (outcome$_i$ > 0, Eq. 4) and the *N*-system only tracks punishing (i.e., aversive) outcomes (outcome$_i$ < 0, Eq. 5); both systems encode the opposite-valence outcomes and null outcomes as though no outcome occurred.

Thus, For the *P*-system, the reward-oriented TD prediction error is

$$\delta_i^P = \begin{cases} \text{outcome}_i + \gamma^{P*} \max_a Q^P(s_{i+1}, \tilde{a}) - Q^P(s_i, a_i) \text{ if outcome}_i > 0 \\ 0 + \gamma^{P*} \max_a Q^P(s_{i+1}, \tilde{a}) - Q^P(s_i, a_i) \text{ if outcome}_i \leq 0 \end{cases} \tag{4}$$

where $0 < \gamma^P < 1$ is the *P*-system temporal discounting parameter (as in TDRL).

The *N*-system similarly encodes a punishment-oriented TD prediction error term

$$\delta_i^N = \begin{cases} \mid \text{outcome}_i \mid + \gamma^{N*} \max_a Q^N(s_{i+1}, \tilde{a}) - Q^N(s_i, a_i) \text{ if outcome}_i < 0 \\ 0 + \gamma^N * \max_a Q^N(s_{i+1}, \tilde{a}) - Q^N(s_i, a_i) \text{ if outcome}_i \geq 0 \end{cases} \tag{5}$$

where $0 < \gamma^N < 1$ is the *N*-system temporal discounting parameter and |outcome$_i$| indicates the absolute value of the (punishing) outcome. We use the absolute value of the outcome so that the *N*-system positively communicates punishments of varying magnitudes, reflecting a neural system that increases its firing rate for larger-than-expected punishments and decreases its firing rate for smaller-than-expected punishments.

The *P*- and *N*-systems prediction errors update expectations of future rewards or punishments of an action, respectively, according to the standard TD learning update rule but for each system independently

$$Q^P(s_i, a_i) \leftarrow Q^P(s_i, a_i) + \alpha^P \delta_i^P \tag{6}$$

$$Q^N(s_i, a_i) \leftarrow Q^N(s_i, a_i) + \alpha^N \delta_i^N \tag{7}$$

where $0 < \alpha^P < 1$ and $0 < \alpha^N < 1$ are learning rates for the *P*- and *N*-systems, respectively; $Q^P(s_i, a_i)$ is the expected state-action value learned by the *P*-system; and $Q^N(s_i, a_i)$ is the expected state-action value learned by the *N*-system.

We compute a composite state-action value for each action by contrasting the *P*- and *N*-system Q values

$$Q(s_i, a_i) \leftarrow Q^P(s_i, a_i) - Q^N(s_i, a_i) \tag{8}$$

which is entered into the categorical logistic choice model (e.g., softmax policy; Eq. 3) as for the TDRL model above.

Alternative reinforcement learning models

Apart from the TDRL and VPRL models described above, we fit "asymmetric" versions of these models to participant choice behavior on the PRP task. Asymmetric TDRL and VPRL models are defined by using distinct learning rate parameters for prediction errors that are positive or negative. For asymmetric TDRL, this amounts to changing Eq. 1 to

$$Q(s_i, a_i) \leftarrow \begin{cases} Q(s_i, a_i) + \alpha^+ \delta_i \text{ if } \delta_i \geq 0 \\ Q(s_i, a_i) + \alpha^- \delta_i \text{ if } \delta_i < 0 \end{cases} \tag{9}$$

where $0 < \alpha^+ < 1$ is the learning rate for positive TD-RPEs and $0 < \alpha^- < 1$ is the learning rate for negative TD-RPEs; the rest of the traditional TDRL model remains the same. For asymmetric VPRL, Eqs. 6 and 7 are changed to

$$Q^P(s_i, a_i) \leftarrow \begin{cases} Q^P(s_i, a_i) + \alpha^{+P} \delta_i^P \text{ if } \delta_i^P \geq 0 \\ Q^P(s_i, a_i) + \alpha^{-P} \delta_i^P \text{ if } \delta_i^P < 0 \end{cases} \tag{10}$$

$$Q^N(s_i, a_i) \leftarrow \begin{cases} Q^N(s_i, a_i) + \alpha^{+N} \delta_i^N \text{ if } \delta_i^N \geq 0 \\ Q^N(s_i, a_i) + \alpha^{-N} \delta_i^N \text{ if } \delta_i^N < 0 \end{cases} \tag{11}$$

where $0 < \alpha^{+P}$ and $\alpha^{-P} < 1$ are learning rate parameters for positive and negative VP-RPEs, respectively, and $0 < \alpha^{+N}$ and $\alpha^{-N} < 1$ are learning rate parameters for positive and negative VP-PPEs, respectively; the rest of the original VPRL model remains the same.

Reinforcement learning hierarchical model parameterization

We specified a hierarchical structure to all computational models to fit participant choice behavior on the PRP task. Individual-level parameter values are drawn from group-level distributions over each model parameter. This hierarchical modeling approach accounts for dependencies between model parameters and biases individual-level parameter estimates toward the group-level mean, thereby increasing reliability in parameter estimates, improving model identifiability and avoiding overfitting (*61*). These hierarchical models therefore cast individual participant parameter values as deviations from a group mean.

Formally, the joint posterior distribution $P(\phi, \theta|y, M_i)$ over group-level parameters ϕ and individual-level parameters θ for the ith model M_i conditioned on the data from the cohort of participants y takes the form

$$P(\mathbf{w} \mid y, M_i) = \frac{p(y \mid \boldsymbol{w}, M_i)p(\boldsymbol{w} \mid M_i)}{p(y \mid M_i)} \tag{12}$$

where we simplify the notation to $P(\mathbf{w}|y, M_i)$, with $\mathbf{w} = \{\phi, \theta\}$) being a parameter vector consisting of all group- and individual-level model parameters for model M_i. Here, $P(y|\mathbf{w}, M_i)$ is the likelihood of choice data y conditioned on the model parameters and hyperparameters, $P(y|M_i)$ is the marginal likelihood (model evidence) of the data given a model, and $P(\mathbf{w}|M_i)$ is the joint prior distribution over model parameters as defined by the model M_i, which can be further factorized into the product of the prior on individual-level model parameters conditioned on the model hyperparameters, $P(\theta|\phi, M_i)$, times the prior over hyperparameters $P(\phi|M_i)$. We define the prior distributions for individual-level model parameters (e.g., $\theta_{\mathrm{TDRL}} = \{\alpha, \tau, \gamma\}$ for M_i = TDRL) and the hyperpriors of the means $-\infty < \mu_{(.)} < +\infty$ and SDs $0 < \sigma_{(.)} < +\infty$ of the population-level parameter distributions (e.g., $\phi_{\mathrm{TDRL}} = \{\mu_\alpha, \mu_\tau, \mu_\gamma, \sigma_\alpha, \sigma_\tau, \sigma_\gamma\}$) to be standard normal distributions. We estimated all parameters in unconstrained space (i.e., $-\infty < \mu_\gamma < +\infty$) and used the inverse Probit transform to map bounded parameters from unconstrained space to the unit interval [0,1] before scaling parameter estimates by the parameter's upper bound

$$\mu_\gamma \sim \mathrm{Normal}(0, 1) \tag{13}$$

$$\sigma_\gamma \sim \mathrm{Normal}^+(0, 1) \tag{14}$$

$$\boldsymbol{\tau}' \sim \mathrm{Normal}(0, 1) \tag{15}$$

$$\boldsymbol{\tau} = \mathrm{Probit}^{-1}(\mu_\gamma + \sigma_\gamma{}^*\boldsymbol{\tau}') * 20 \tag{16}$$

where bold terms indicate a vector of parameter values over participants. This noncentered parameterization (*62*) and inverse Probit transformation creates a uniform prior distribution over individual-level model parameters between specified lower and upper bounds. Note that, for learning rate and temporal discount parameters, the scaling factor (upper bound) was set to 1, whereas it was set to 20 for the choice temperature parameter. We used the Hamiltonian Monte Carlo sampling algorithm in the probabilistic programming language Stan (*63*) via the R package rstan (v. 2.21.2) to sample the joint posterior distribution over group- and individual-level model parameters for both cohorts individually and for all participants combined into a single cohort. For all models and each cohort, we executed 12,000 total iterations (2000 warm-up) on each of three chains for a total of 30,000 posterior samples per model parameter. We inspected chains for convergence by verifying sufficient chain mixing according to the Gelman-Rubin statistic $\hat{R}$, which was less than 1.1 for all parameters.

Reinforcement learning model comparison

We compared the fit of each model to participant choice behavior on the PRP task according to their model evidence (i.e., Bayesian marginal likelihood), which represents the probability or "plausibility" of observing the actual PRP task data under each model (*64*). In Bayesian model comparison, the model with the greatest posterior model probability $p(M_i \mid y)$ is deemed the best explanation for the data y and is computed by

$$P(M_i \mid y) \propto P(y \mid M_i)P(M_i) \tag{17}$$

where $P(y|M_i)$ is the model marginal likelihood (i.e., "model evidence"), the normalizing constant of Eq. 12, and $P(M_i)$ is the model's prior probability. The model evidence is defined as

$$P(y \mid M_i) = \int P(y \mid \mathbf{w}, M_i)P(\mathbf{w} \mid M_i)d\mathbf{w} \tag{18}$$

where $P(\mathbf{w}|M_i)$ is the prior probability of a model M_i's parameters $\mathbf{w}$ before observing any data and $P(y|\mathbf{w}, M_i)$ is the likelihood of data y given a model and its parameters.

The marginal likelihood for each model as defined in Eq. 18 is an optimal measure for performing model comparison as it represents the balance between the fit of each model to the cohort's data (as captured by the first term in the integral) and the complexity of each model (as captured in the second term of the integral), integrated over all sampled model parameter values. In effect, although more complex or flexible models (i.e., more parameters) are able to predict a greater variety of behaviors and therefore increase the data likelihood $P(y|\mathbf{w}, M_i)$, more complex models have a higher dimensional parameter space and therefore must necessarily assign lower prior probability to the parameter values $P(\mathbf{w}|M_i)$. In this way, the marginal likelihood of a model automatically penalizes model complexity, sometimes referred to as the "Bayesian Occam razor" (*64*).

To compare the TDRL and VPRL models (i.e., M_1 and M_2, respectively), the relative posterior model probability can be defined as

$$\frac{P(M_1 \mid y)}{P(M_2 \mid y)} = \frac{P(M_1) * P(y \mid M_1)}{P(M_2) * P(y \mid M_2)} \tag{19}$$

where the ratio of posterior model probabilities $P(M_1 \mid y)/P(M_2 \mid y)$ is referred to as the "posterior odds" of TDRL relative to VPRL; $P(M_1)$ and $P(M_2)$ are the prior probabilities of the TDRL and VPRL models, respectively; and the ratio of marginal likelihoods $P(y \mid M_1)/P(y \mid M_2)$ is termed the "Bayes factor," which is a standard measure for Bayesian model comparison. By assigning equal prior probabilities over the set of candidate models, each model's evidence $P(y|M_i)$ can be used to rank each model in the set for comparison. The marginal likelihoods are computed as log-scaled, and, therefore, the Bayes factor is computed as the difference between log marginal likelihoods for two models; a positive value for the Bayes factor indicates greater support for M_1 (the model in the numerator of Eq. 19), whereas a negative value for the Bayes factor indicates greater support for M_2. We estimated the log model evidence (marginal likelihood) for all models for each cohort and for all participants combined into a single cohort using an adaptive importance sampling routing called bridge sampling as implemented in the R package bridgesampling [v. 1.1-2; (*65*)]. Bridge sampling is an efficient and accurate approach to calculating normalizing constants like the marginal likelihood of models even with hierarchical structure and for reinforcement learning models in particular (*65*). To further ensure stability in the bridge sampler's estimates of model evidence, we performed 10 repetitions of the sampler and report the median and interquartile range of the estimates of model evidence. The model with the maximum (i.e., least negative) model evidence is the preferred model.

In addition to the standard Bayesian model comparison using model marginal likelihoods, we estimated each model's Bayesian leave-one-out (LOO) cross-validation predictive accuracy, defined as a model's expected log predictive density [ELPD-LOO; (*66*)]

$$\mathrm{elpd}_{\mathrm{LOO}} = \sum_{i=1}^{N} \log(p(y_i \mid y_{-i})) \tag{20}$$

where the posterior predictive distribution $p(y_i|y_{-i})$ for held-out data y_i, given a set of training data y_{-i}, is

$$P(y_i \mid y_{-i}) = \int p(y_i \mid \mathbf{w}) p(\mathbf{w} \mid y_{-i}) d\mathbf{w} \tag{21}$$

The ELPD is an estimate of (i.e., approximation to) the cross-validated accuracy of a given model in predicting unobserved (i.e., held-out) participant data, given the posterior distribution over model parameters fit to a training set of participant data (*66*). We approximate this integral via importance sampling of the joint posterior parameter distribution given the training data $p(\mathbf{w}|y_{-i})$ using the R package loo [v. 2.3.1; (*66*)].

We repeated this model comparison analysis (table S1) for the behavior-only cohort and a "meta-analytic" cohort combining the patients with ET and behavioral participants ($N = 53$). Running the model comparison analysis in triplicate allowed us to assess the replicability of the model comparison results, and using multiple model comparison criteria allowed us to assess the robustness and generalizability of the model comparison results. We elected to focus the subsequent behavioral and neurochemical analyses on the basic TDRL and VPRL models because the computational differences between these models most directly address the neurobiological mechanism that was our main target of investigation: the partitioned signaling of RPEs and PPEs; all subsequent behavioral analyses and neurochemical time series analyses of the ET cohort used the computational model fits to the ET cohort alone.

Model and parameter recovery

We performed a model recovery analysis to validate that our Bayesian model comparison analysis is able to accurately identify the true generative model of choice behavior on the PRP task. For this model recovery analysis, we simulated choice behavior on the PRP task for both the ET ($N = 11$) and behavioral ($N = 42$) cohorts using the mean individual-level parameter values for TDRL and VPRL models and then computed model comparison criteria for the TDRL and VPRL models to determine whether the model comparison analysis identified the true generative model as the best model (table S2).

To validate that our hierarchical computational model fitting procedure is able to accurately estimate model parameters for each participant and for TDRL and VPRL models, we performed a parameter recovery analysis. We determined whether the empirical parameter distributions for both cohorts were credibly different by computing the difference between the ET and behavioral cohorts' group-level TDRL and VPRL parameter distributions, which revealed no credible differences in any TDRL or VPRL model parameter between the cohorts (fig. S9). Given this result and because the larger sample size in the behavioral cohort increases the robustness of the parameter recovery analysis results, we elected to perform the parameter recovery analysis using the behavioral cohort's data. We first calculated the mean TDRL and VPRL parameter values for each participant in the behavioral cohort to simulate choice datasets ($N = 42$) on the PRP task (using different option presentation sequences), refitted the TDRL and VPRL models to the simulated PRP dataset, and then computed the Pearson's correlation coefficient between the mean model parameters fitted to the actual participant PRP data and the simulated PRP data.

Electrochemistry data analysis

General description of human voltammetry approach

The human FSCV protocol used in the current study has been extensively described in previous publications (*40*–*43*), and, therefore, we give a brief general description here. The human voltammetry protocol, which involves the construction of custom carbon-fiber microelectrodes for use in the human brain (*40*, *42*), is designed as a human-level extension of traditional voltammetry protocols used in model organism (e.g., rodent) and ex vivo slice or culture preparations. The specific electrochemical properties of the custom electrodes used in the human voltammetry protocol have been validated in the rodent brain as matching those of rodent electrodes (*42*). In addition, the voltage waveform and cycling frequency of the stimulating current, as well as the sampling rate of the current time series during the voltage sweeps used in the human protocol, are identical to those used in rodent studies (*27*).

The central difference between the human voltammetry protocol used here (*40*, *41*, *43*) and traditional voltammetry protocols is the statistical method used to estimate the in vivo concentration of different neurochemical analytes. Specifically, in traditional voltammetry protocols, estimating the concentration of an analyte of interest (e.g., dopamine) involves performing principal components regression on recorded currents (voltammograms), wherein the principal component time series used as regressors are derived from an in vitro dataset of voltammograms of known concentrations of the analyte of interest. By contradistinction, the statistical method used for analyte concentration estimation in the human voltammetry protocol adopts a supervised statistical learning approach. This approach involves training an elastic net-penalized linear regression model on in vitro voltammograms of known concentrations of analytes of interest (e.g., dopamine and serotonin), varying levels of pH, and common metabolites of target analytes [e.g., 3,4-dihydroxyphenylacetic acid (DOPAC) and 5-hydroxyindoleacetic acid (5-HTIAA)] or other neurotransmitters [e.g., norepinephrine; (*67*)]. In this protocol, multiple carbon-fiber microelectrodes identical to those used for human recordings were used to collect the in vitro training datasets, and the penalized linear regression model is optimized via cross-validation to reduce the out-of-probe error. This penalized cross-validation procedure has the added benefits of reducing bias in model performance due to overfitting on training data and automatically selecting and regularizing model coefficient values (via the elastic net), thereby providing reliable estimation performance when recovering analyte concentrations from the electrodes used during the human voltammetry experiments. This approach provides more reliable estimates of dopamine than principal components regression (*40*), especially under different pH levels. In addition, this approach reliably and accurately differentiates mixtures of dopamine and serotonin from a background of varying pH (*41*, *42*) and changing levels of dopamine or serotonin metabolites or other neurochemicals like norepinephrine (*67*).

FSCV carbon-fiber microelectrodes and experimental protocol

The FSCV protocol as well as the construction of carbon-fiber microelectrode probes and the specifications of the mobile electrochemistry recording station have been extensively described in previous work (*40*, *42*). Briefly, custom carbon-fiber microelectrodes for human FSCV experiments were placed in the caudate nucleus as determined by DBS surgery planning for patients with ET. We note that electrode placement within the caudate nucleus is different for each patient in accordance with the patient-specific trajectory of the DBS electrode used for treatment. The FSCV protocol consisted of an equilibration phase and an experiment phase where the voltammetry measurement waveform—a triangular waveform starting at −0.6 V, ramping up to a peak of +1.4 V at 400 V/s, and ramping back down to −0.6 V at −400 V/s—was first cycled at 60 Hz for 10 min to allow for equilibration of the electrode surface followed by a 10-Hz application of the waveform for the duration of the experimental window encompassing the behavioral task. All recordings of the measurement waveform-induced currents (voltammograms) were collected at a 100-kHz sampling rate.

In vitro training data protocol and neurochemical concentration estimation model training

The in vitro data collected to train the dopamine concentration estimation model consisted of a population of five carbon-fiber microelectrodes identical to those used in the human voltammetry experiments. Each probe contributed 16 datasets (one per solution mixture), with each dataset consisting of 2 min worth of voltammogram recordings in mixture solutions of known concentrations of dopamine, DOPAC, and ascorbic acid (from 0 to 1500 nM in 100 nM increments), with a background of varying pH levels (from 7.2 to 7.6 in 0.1 increments). All voltammograms in the training datasets were sampled at 250 kHz (resulting in 2500 samples per voltammogram) and then downsampled by averaging every 15 samples. The voltammograms used to train the dopamine concentration estimation model were taken over the last 90 s of a probe's 2-min recording in a given solution, as these later time points are less affected by flow or electrode equilibration artifacts that occur in the beginning of recording periods. Each probe therefore contributed a total of 900 voltammograms per each of 16 solution mixtures resulting in a total of 14,400 labeled samples per probe, each corresponding to the probe's response to mixed levels of dopamine, DOPAC, ascorbic acid, and pH.

Using this in vitro training dataset, we fit a penalized linear regression model using the elastic net algorithm (*68*) to predict known concentrations of each analyte, optimized using 10-fold cross-validation. In this model, the target variable (y) is an N-by-4 matrix of known levels of dopamine, DOPAC, ascorbic acid, and pH, with N = 12,960 samples (9/10ths of the 14,400 total samples, with 1/10 held-out for cross-validation); the predictor variable matrix (x) is an N-by-498 matrix of the corresponding raw and differentiated voltammograms (167 time points per down-sampled voltammogram, plus 166 time points for its first derivative and 165 time points for its second derivative). The linear model coefficients (β) are determined by minimizing the residual sum of squares, subject to the elastic net penalty

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^{\mathrm{T}}\beta\right)^2 + \lambda P_\alpha(\beta) \qquad (22)$$

where λ is a penalty term that weighs the influence of the elastic net penalty, $P_\alpha(\beta)$

$$P_\alpha(\beta) = (1-\beta)\frac{1}{2}\|\beta\|_{\ell_2}^2 + \alpha\|\beta\|_{\ell_1} \qquad (23)$$

where $0 < \alpha < 1$ parameterizes the relative weighting between the ridge (ℓ_2-norm) and lasso (ℓ_1-norm) regularizations. The optimal values of β, λ, and α are determined using a 10-fold cross-validation procedure via the cvglmnet function of the glmnet package in MATLAB. Here, we fixed α = 1 and used the smallest λ value to estimate dopamine concentrations from in vivo experimental recordings.

Dopamine time series analysis

Time series of dopamine concentrations for each participant were generated from the optimized elastic net linear regression model with 100-ms temporal resolution. We first cut out individual trials' time series from 1 s (10 samples) before the trial's option presentation screen to 100 ms (1 sample) before the next trial's option presentation, z-scored the dopamine concentrations within each trial, and smoothed the within-trial dopamine time series using a 0.3 s (3 samples) sliding-window lagging average (*43*). From these individual trial time series, we extracted individual event-related dopamine responses lasting from 300 ms before event to 700 ms (i.e., 11 time points total) following option presentation, action selection, and outcome presentation. We normalized each event-related time series by subtracting the mean dopamine level in the 300- to 0-ms window leading up to the onset of each event (i.e., four samples) and then dividing by the SD in dopamine levels in the 300- to 0-ms baseline window. This normalization produces three matrices (size trials × time) of baseline-corrected dopamine time series, one for each event across all trials. The event-related dopamine time series were then sorted according to whether TDRL or VPRL models specified the time series as positive or negative RPEs or PPEs. This process was repeated for each patient included in the voltammetry analysis (n = 3).

For group-level analyses, we either performed a fixed-effects analysis, wherein we grouped dopamine responses across all events, trials, and patients before conducting statistical comparisons of prediction error responses or instead performed a mixed-effects analysis where we first averaged dopamine prediction error responses over all events and trials for each patient individually and then performed statistical comparisons using the three patients' mean prediction error responses. Parametric statistical testing consisted of performing either two-way ANOVA tests (prediction error sign, time) of dopamine prediction error responses (Fig. 2) or independent samples t tests at single time points to compare dopamine responses to positive and negative RPEs and PPEs (Fig. 2). Nonparametric statistical testing (fig. S4) consisted of conducting permutation tests (50,000 iterations) where we computed the difference between the mean dopamine response to positive and negative RPEs and PPEs (permuted labels) at each time point, divided this difference measure by the summed variance in dopamine levels at that time point, and computed *P* values as the percentage of permuted mean difference measures that were greater than the absolute value of the actual mean difference. Correction for multiple comparisons was performed using the Benjamini-Hochburg procedure using a false discovery rate level α = 0.05.

Dopamine prediction error ROC decoding analysis

For the receiver operating characteristic (ROC) analysis (Fig. 3), we trained logistic regression models on segments of event-related dopamine fluctuations to classify positive and negative RPEs and PPEs. We trained separate classifiers using either TDRL or VPRL computational model-defined fluctuations; that is, the event-related dopamine signals used to train each classifier differed according to whether TDRL and VPRL models specified an event as being either a positive or negative RPE or PPE. For the RPE classifiers, we trained the logistic models for TDRL and VPRL using samples from 0 to 300 ms of the dopamine fluctuations; for the PPE classifiers, we used samples from 400 to 700 ms of the dopamine fluctuations. These RPE- and PPE-specific temporal windows were chosen on the basis of our findings from the dopamine time series analysis (Fig. 2). From the fitted classifiers, we computed the area under the ROC curve (auROC) separately for the TDRL- and VPRL-based classifiers using the perfcurve function in MATLAB. We compared the relative performance of the TDRL and VPRL classifiers for decoding positive and negative RPEs and PPEs using a permutation test where we computed the difference in auROC values across 10,000 iterations and compared the true auROC values to the permutation test null distribution to obtain *P* values. *P* values were computed as the number of permutation test samples that were greater than the true difference in TDRL and VPRL classifier auROC values.

Note that, for the prediction error classifier analyses, only 68% of positive TD-RPEs (429 events) were classified as positive VP-RPEs (i.e., only 68% agreement between the TDRL and VPRL classifiers on positive RPE signals), with 26% being classified as negative VP-PPEs (168 events), 4% as positive VP-PPEs (25 events), and 2% as negative VP-RPEs (18 events). Similarly, for negative TD-RPEs, only 45% are classified as negative VP-RPEs (234 events), whereas 38% are classified as positive VP-PPEs (204 events), 14% as negative VP-PPEs (77 events), and 2% as positive VP-RPEs (9 events).

Supplementary Materials

This PDF file includes:
Figs. S1 to S9
Tables S1 and S2

REFERENCES AND NOTES

1. P. R. Montague, S. E. Hyman, J. D. Cohen, Computational roles for dopamine in behavioural control. *Nature* **431**, 760–767 (2004).
2. D. J. Moore, A. B. West, V. L. Dawson, T. M. Dawson, Molecular pathophysiology of Parkinson's disease. *Annu Rev Neurosci.* **28**, 57–87 (2005).
3. Q. J. M. Huys, N. D. Daw, P. Dayan, Depression: A decision-theoretic analysis. *Annu Rev Neurosci.* **38**, 1–23 (2015).
4. D. Redish, Addiction as a computational process gone awry. *Science* **306**, 1944–1947 (2004).
5. D. Redish, J. Gordon, Eds., *Computational Psychiatry: New Perspectives on Mental Illness* (MIT Press, 2016).
6. S. N. Haber, The primate basal ganglia: Parallel and integrative networks. *J Chem Neuroanat.* **26**, 317–330 (2013).
7. P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
8. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
9. H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
10. H. M. Bayer, B. Lau, P. W. Glimcher, Statistics of midbrain dopamine neuron spike trains in the awake primate. *J. Neurophysiol.* **98**, 1428–1439 (2007).
11. P. W. Glimcher, Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proc Natl Acad Sci U S A.* **108**, 15647–15654 (2011).
12. A. S. Hart, R. B. Rutledge, P. W. Glimcher, P. E. M. Phillips, Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J. Neurosci.* **34**, 698–704 (2014).
13. N. Eshel, J. Tian, M. Buckwich, N. Uchida, Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
14. R. Amo, S. Matias, A. Yamanaka, K. F. Tanaka, N. Uchida, M. Watabe-Uchida, A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nat. Neurosci.* **25**, 1082–1092 (2022).
15. R. S. Sutton, A. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998), 9, 1054.
16. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Reidmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassibis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
17. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Grapel, D. Hassabis, Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
18. O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
19. P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, H. Kitano, Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602**, 223–228 (2022).
20. K. A. Zaghloul, J. A. Blanco, C. T. Weidemann, K. McGill, J. L. Jaggi, G. H. Baltuch, M. J. Kahana, Human substantia nigra neurons encode unexpected financial rewards. *Science* **323**, 1496–1499 (2009).
21. S. M. McClure, G. S. Berns, P. R. Montague, Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
22. J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, R. J. Dolan, Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
23. M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan, C. D. Frith, Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042–1045 (2006).
24. T. Patriarchi, J. R. Cho, K. Merten, M. W. Howe, A. Marley, W.-H. Xiong, R. W. Folk, G. J. Broussard, R. Liang, M. J. Jang, H. Zhong, D. Dombeck, M. V. Zastrow, A. Nimmerjahn, V. Gradinaru, J. T. Williams, L. Tian, Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science* **360**, eaat4422 (2018).
25. F. Sun, J. Zhou, B. Dai, T. Qian, J. Zeng, X. Li, Y. Zhou, Y. Zhang, Y. Wang, C. Qian, K. Tan, J. Feng, H. Dong, D. Lin, G. Cui, Y. Li, Next-generation GRAB sensors for monitoring dopaminergic activity in vivo. *Nat. Methods* **17**, 1156–1166 (2020).
26. M. G. Kutlu, J. E. Zachry, P. R. Meulgin, S. A. Cajigas, M. F. Chevee, S. J. Kelly, B. Kutlu, L. Tian, C. A. Siciliano, E. S. Calipari, Dopamine release in the nucleus accumbens core signals perceived saliency. *Curr. Biol.* **31**, 4748–4761.e8 (2021).
27. P. E. M. Phillips, G. D. Stuber, M. L. A. V. Heien, R. M. Wightman, R. M. Carelli, Subsecond dopamine release promotes cocaine seeking. *Nature* **422**, 614–618 (2003).
28. I. Willuhn, L. M. Burgeno, B. J. Everett, P. E. M. Philliips, Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proc Natl Acad Sci U S A* **109**, 20703–20708 (2012).

29. A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M. Vander Weele, R. T. Kennedy, B. J. Aragona, J. D. Burke, Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
30. J. Mirenowicz, W. Schultz, Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* **379**, 449–451 (1996).
31. M. Matsumoto, O. Hikosaka, Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).
32. J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, N. Uchida, Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
33. C. D. Fiorillo, Two dimensions of value: Dopamine neurons represent reward but not aversiveness. *Science* **341**, 546–549 (2013).
34. C. D. Fiorillo, M. R. Song, S. R. Yun, Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. *J. Neurosci.* **33**, 4710–4725 (2013).
35. H. Matusmoto, J. Tian, N. Uchida, M. Watabe-Uchida, Midbrain dopamine neurons signal aversion in a reward-context dependent manner. *eLife* **5**, e17328 (2016).
36. B. Seymour, J. P. O'Doherty, P. Dayan, M. Koltzenburg, A. K. Jones, R. J. Dolan, K. J. Friston, R. S. Frackowiak, Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667 (2004).
37. B. Seymour, J. P. O'Doherty, M. Koltzenburg, K. Wiech, R. Frackowiak, K. Friston, R. Dolan, Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat. Neurosci.* **8**, 1234–1240 (2005).
38. B. Seymour, N. Daw, P. Dayan, T. Singer, R. Dolan, Differential encoding of losses and gains in the human striatum. *J. Neurosci.* **27**, 4826–4831 (2007).
39. M. R. Delgado, J. Li, D. Schiller, E. A. Phelps, The role of the striatum in aversive learning and aversive prediction errors. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **363**, 3787–3800 (2008).
40. K. T. Kishida, I. Saez, T. Lohrenz, M. R. Witcher, A. W. Laxton, S. B. Tatter, J. P. White, T. L. Ellis, P. E. M. Phillips, P. R. Montague, Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proc Natl Acad Sci U S A.* **113**, 200–205 (2016).
41. R. J. Moran, K. T. Kishida, T. Lohrenz, I. Saez, A. W. Laxton, M. R. Witcher, S. B. Tatter, T. L. Ellis, P. E. M. Phillips, P. Dayan, P. R. Montague, The protective action encoding of serotonin transients in the human brain. *Neuropsychopharmacology* **43**, 1425–1435 (2018).
42. K. T. Kishida, S. G. Sandberg, T. Lohrenz, Y. G. Comair, I. Sáez, P. E. M. Phillips, P. R. Montague, Sub-second dopamine detection in human striatum. *PLOS ONE* **6**, e23291 (2011).
43. D. Bang, K. T. Kishida, T. Lohrenz, J. P. White, A. W. Laxton, S. B. Tatter, S. M. Fleming, P. R. Montague, Sub-second dopamine and serotonin signaling in human striatum during perceptual decision-making. *Neuron* **108**, 999–1010.e6 (2020).
44. P. R. Montague, K. T. Kishida, R. J. Moran, T. M. Lohrenz, An efficiency framework for valence processing systems inspired by soft cross-wiring. *Curr. Opin. Behav. Sci.* **11**, 121–129 (2016).
45. K. T. Kishida, L. P. Sands, "A dynamic affective core to bind the contents, context, and value of conscious experience" in *Affect Dynamics*, C. Waugh, P. Kuppens, Eds. (Springer, 2021), pp. 293–328.
46. L. P. Sands, A. Jiang, R. E. Jones, J. D. Trattner, K. T. Kishida, Valence-partitioned learning signals drive choice behavior and phenomenal subjective experience in humans. biorxiv 2023.03.17.533213 [**Preprint**]. 2023. https://doi.org/10.1101/2023.03.17.533213.
47. D. Haubenberger, M. Hallett, Essential tremor. *N Engl J Med.* **378**, 1802–1810 (2018).
48. B. Liebenow, M. Williams, T. Wilson, I. U. Haq, M. S. Siddiqui, A. W. Laxton, S. B. Tatter, K. T. Kishida, Intracranial approach for sub-second monitoring of neurotransmitters during DBS electrode implantation does not increase infection rate. *PLOS ONE* **17**, e0271348 (2022).
49. R. A. Rescorla, A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory* 2, 64–69 (1972).
50. N. D. Daw, P. Dayan, Opponent interactions between serotonin and dopamine. *Neural Netw.* **15**, 603–616 (2002).
51. A. G. E. Collins, M. J. Frank, Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121**, 337–366 (2014).
52. W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, M. Botvinick, A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
53. C. Liu, P. S. Kaeser, Mechanisms and regulation of dopamine release. *Curr. Opin. Neurobiol.* **57**, 46–53 (2019).
54. H. Jeong, A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, V. M. K. Namboodiri, Mesolimbic dopamine release conveys causal associations. *Science* **378**, eabq6740 (2022).
55. M. Matsumoto, O. Hikosaka, Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447**, 1111–1115 (2007).
56. M. Watabe-Uchida, N. Eshel, N. Uchida, Neural circuitry of reward prediction error. *Annu Rev Neurosci.* **40**, 373–394 (2017).
57. C. Xiao, J. Wei, G. Zhang, C. Tao, J. J. Huang, L. Shen, I. R. Wickersham, H. W. Tao, L. I. Zhang, Glutamatergic and GABAergic neurons in pontine central gray mediate opposing valence-specific behaviors through a global network. *Neuron* **111**, 1486–1503.e7 (2023).
58. Y. Du, S. Zhou, C. Ma, H. Chen, A. Du, G. Deng, Y. Liu, A. J. Tose, L. Sun, Y. Liu, H. Wu, H. Lou, Y. Yu, T. Zhao, S. Lammel, S. Duan, H. Yang, Dopamine release and negative valence gated by inhibitory neurons in the laterodorsal tegmental nucleus. *Neuron* **111**, 3102–3118.e7 (2023).
59. S. Lammel, B. K. Lim, C. Ran, K. W. Huang, M. J. Betley, K. M. Tye, K. Deisseroth, R. C. Malenka, Input-specific control of reward and aversion in the ventral tegmental area. *Nature* **491**, 212–217 (2012).
60. C. J. C. H. Watkins, P. Dayan, Q-learning. *Machine Learning* **8**, 279–292 (1992).
61. W.-Y. Ahn, N. Haines, L. Zhang, Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Comput Psychiatr.* **1**, 24–57 (2020).
62. O. Papaspiliopoulos, G. O. Roberts, M. Sköld, A general framework for the parametrization of hierarchical models. *Statistical Science* **22**, 59–73 (2007).
63. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
64. D. J. McKay, *Information Theory, Inference, and Learning Algorithms*. (Cambridge University Press, 2003).
65. Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, H. Steingroever, A tutorial on bridge sampling. *J Math Psychol.* **81**, 80–97 (2017).
66. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432 (2017).
67. P. R. Montague, K. T. Kishida, Computational underpinnings of neuromodulation in humans. *Cold Spring Harb. Symp. Quant. Biol.* **83**, 1425–1435 (2018).
68. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B. Stat. Methodo.* **67**, 301–320 (2005).

Acknowledgments: We thank the participants of this study and the research and surgical nursing staff at Atrium Health Wake Forest School of Medicine for support and cooperation. **Funding:** This work was supported by National Institutes of Health grants R01MH121099 (K.T.K.), R01DA048096 (K.T.K.), R01MH124115 (K.T.K.), P50DA006634 (K.T.K.), 5KL2TR001420 (K.T.K.), F31DA053174 (L.P.S.), T32DA041349 (L.P.S.), and F30DA053176 (B.L.). **Author contributions:** L.P.S. designed and performed behavioral and FSCV data analysis, interpreted results, wrote and edited original manuscript drafts, and approved final manuscript. A.J. coded behavioral tasks, collected data, analyzed FSCV data, and edited and approved final manuscript. B.L. collected data and edited and approved final manuscript. E.D. collected data and edited and approved final manuscript. A.W.L. conceived of surgical strategies for safe and effective placement of electrodes for human FSCV, performed surgical placement of electrodes for human FSCV experiment, collected data, and edited and approved final manuscript. S.B.T. conceived of surgical strategies for safe and effective placement of electrodes for human FSCV, performed surgical placement of electrodes for human FSCV experiment, collected data, and edited and approved final manuscript. P.R.M. interpreted results and edited and approved final manuscript. K.T.K. conceived the study, designed experiments, supervised and guided data collection and analysis, interpreted results, wrote and edited original manuscript drafts, and approved final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Anonymized individual-level participant behavioral task data and neurochemical time series data used in this study may be made available upon submission of a formal project outline from any qualified investigator to the corresponding author and subsequent approval by the corresponding author in line with data protection regulations of Wake Forest University School of Medicine IRB. Custom-written analysis scripts for generating the behavioral and neurochemical results of this manuscript can be found at https://zenodo.org/doi/10.5281/zenodo.10067903.

Submitted 28 April 2023
Accepted 3 November 2023
Published 1 December 2023
10.1126/sciadv.adi4927