Title

Striatal Gradient in Value-Decay Explains Regional Differences in Dopamine Patterns and Reinforcement Learning Computations

Abbreviated title

Value-decay links dopamine patterns and functions

Authors

Ayaka Kato[1,2] & Kenji Morita[3,4*]

[1] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674, United States

[2] Postdoctral Fellowship for Research Abroad, Japan Society for the Promotion of Science, Tokyo 102-0083, Japan

[3] Physical and Health Education, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan

[4] International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo, Tokyo 113-0033, Japan

Corresponding author

[*] Kenji Morita, Ph.D. (ORCID iD: orcid.org/0000-0003-2192-4248)

Physical and Health Education, Graduate School of Education, The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

morita@p.u-tokyo.ac.jp

Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgments

KM was supported by Grants-in-Aid for Scientific Research 23H03295 and 23K27985 from Japan Society for the Promotion of Science (JSPS) and the Naito Foundation. AK was supported by JSPS Overseas Research Fellowships.

Abstract

Dopamine has been suggested to encode reward-prediction-error (RPE) in reinforcement learning (RL) theory, but also shown to exhibit heterogeneous patterns depending on regions and conditions: some exhibiting ramping response to predictable reward while others only responding to reward-predicting cue, and some lacking response to unpredictable reward. It remains elusive how these heterogeneities relate to various RL algorithms suggested to be employed by animals/humans, such as RL under predictive state representation, hierarchical RL, and distributional RL. Here we show that these relations could be coherently explained if the decay of learned values (value-decay), implemented by the decay of dopamine-dependent plastic changes in the synaptic strengths, is considered. First, we show that value-decay causes ramping RPE under traditional non-predictive representations but not under the predictive successor representation (SR). This explained the observed gradual fading of dopamine ramping over repeated reward navigation by gradual formation of SR. Next, we constructed a hierarchical RL model that coupled two systems with and without value-decay. The model explained the observed distinct patterns of neuronal activities in parallel striatal-dopamine circuits and their suggested distinct functions in flexible learning versus stable habit. Lastly, we examined two distinct algorithms of distributional RL with and without value-decay. These algorithms explained how the reported distinct dopamine patterns in different striatal regions are linked to the suggested different strengths of distributional coding in these regions. These results suggest that within-striatum differences, or more specifically, medial-lateral gradient in value/synaptic-decay tunes regionally different RL computations through generating distinct patterns of DA/RPE.

Introduction

Dopamine (DA) is crucially involved in value learning and motivated behavior. In reward learning task, it was originally found that DA shows an abrupt response to reward-predictive cue but not to reward itself after learning, and such a pattern was suggested to represent the temporal-difference reward-prediction-error (TD-RPE) defined in the reinforcement learning (RL) theory [1,2]. Subsequent studies found different DA patterns, in particular, ramping DA signals towards reward, which sustained after learning [3-6] although gradual fading was also observed [7]. Ramping DA signals were demonstrated to encode TD-RPE by experiments using virtual reality [5], while other studies suggested that ramping DA encodes state value [4,6].

Theoretically, there are multiple ways of how TD-RPE can entail sustained/ramping patterns [5,8-12]. One of them is the decay of learned values (hereafter referred to as value-decay), or biologically, the decay of DA-dependent plastic changes in the strength of corticostriatal synapses [9,11] that stores value memory [13-15]. A recent model exhibiting ramping TD-RPE [16] also in effect assumed value-decay. Notably, according to the formulae derived for RL with value-decay [9], ramping TD-RPE was proportional to state value except for cue response, and so the TD-RPE and state-value accounts of ramping DA could be reconciled given the presence of value-decay. While roles of decay/forgetting of memories have been actively studied [17-20], value-decay is not assumed in standard RL [21], but was shown to be beneficial for adaptation to stochastic or changing environments [22,23], self-paced behavior [11], and energy reserve [24] and also potentially better fit behavior [25,26]. Moreover, recent studies [27-29] suggest that the striatum-DA system implements distributional RL [30-33], and value-decay was assumed to ensure stability against value divergence in certain distributional RL models [27,28] while not in other models [29,30]. Value-decay was also suggested to be useful for canceling out uncertainty-induced bias in value learning [12] if correctly tuned [34].

While recent studies revealed mechanisms for forgetting/decay of memories [18,35-38], including those depending on DA, in Drosophila [36,39] and mammals [40,41], so far there is no direct demonstration of the decay of the strength of cortico-striatal synapses that stores value memory [13-15] at the best of our knowledge (although possible occurrence of DA-dependent forgetting in the nucleus accumbens has been suggested [42]). However, there were findings that appear to suggest the existence of value-decay. Specifically, monkey caudate head (CDh) neurons rapidly developed a response encoding cue value within a session but lost it at the beginning of sessions in subsequent days, while caudate tail (CDt) neurons hardly developed value response within a session but gradually developed it across days [43] (Figure 6B,C of [43]). DA neurons in rostral-medial and caudal-lateral substantia nigra pars compacta (rm/clSNc), which project to CDh and CDt, respectively, also lost and retained predicted-value response [44]. These results appear to suggest that CDh-rmSNc circuit entails value-decay whereas CDt-clSNc circuit does not.

An emerging possibility is that the presence or absence of value-decay could explain how different patterns of DA signals shown in different regions or conditions serve for distinct computations operated in those regions/conditions. Although it was originally considered that the DA-cortico-striatal circuits implement simple RL algorithms, later studies suggested that diverse sophisticated RL algorithms may be implemented in different regions or under different conditions. There are (at least) three sorts of sophisticated RL algorithms that were suggested to be implemented: (i) model-based RL [45-47] and/or RL using sophisticated state

representations, in particular, predictive representations such as the successor representation (SR) [48], which enables partially model-based-like (e.g., outcome devaluation-sensitive) behavior [49-51], (ii) hierarchical RL, in which linked parallel systems learn to achieve hierarchical behavioral goals [52-57], and (iii) distributional RL, which learns not only the mean but also the distribution of outcomes [31,32]. We examine how value-decay in these RL algorithms can explain the different patterns of DA signals and their functional relevance.

Results

Value-decay in RL under different state representations

We examined the pattern of TD-RPE generated in RL models with value-decay under different state representations in a simulated cue-reward Pavlovian association learning task, in which a single cue was followed by a single reward of a fixed amount after a fixed delay (see the Methods for the details of the models and the simulations). First we examined a case where discretized individual timings from cue to reward were represented separately as punctate states (Fig. 1Aa), as done previously [9]. In this case, if there was value-decay, TD-RPE showed a sustained ramp towards reward after learning, in addition to an abrupt response to cue (Fig. 1Ac). As the rate of value-decay increased, ramping became more prominent, while the abrupt response to cue degraded (Fig. 1Ad).

Next we examined cases where the states were represented by features. Specifically, we assumed that each state was represented by a high-dimensional feature vector, and state values were calculated as weighted sums of features with subtraction of constants. The weights were updated according to TD-RPE multiplied with features [21]. Biologically, features were assumed to be encoded by cortical neurons and the update was assumed to be implemented by DA-dependent plasticity of corticostriatal synapses [1,21]. We further assumed non-negativity of each weight for biological plausibility, although negative weights were effectively allowed by the assumed subtraction of constants in the value calculation (presumably implemented by striatal inhibitory interneurons). We also assumed a larger learning rate for positive than negative TD-RPE, in reference to findings and suggestions on the property of D1 receptor-dependent plasticity [27-29,58,59].

Since it has been suggested that cortical representations may be sparse [60], we examined a case where, for each feature vector, only a small proportion (10%) of elements had positive pseudo-random values (Fig. 1Ba). As the rate of decay of value-weights (to the initial weights) increased, TD-RPE exhibited a ramp towards reward (Fig. 1Bc,d), similarly to the case with punctate representation. We also examined a case with dense representation, where all the elements of each feature vector had positive pseudo-random values (Fig. 1Ca). In this case, value-weight-decay still caused ramping of TD-RPE, but its degree was smaller than the case with sparse representation (Fig. 1Cc,d).

We further examined a case where the states were represented by future occupancies of other (and self) states, namely, by the successor representation (SR) [48] (Fig. 1Da), which enables partially model-based-like behavior [49,51] and has been suggested to be actually used by humans [50] and implemented in certain parts of cortico-striatal-DA system [61,62]. In this case, value-weight-decay did not cause a ramp of TD-RPE while still causing sustained response to reward (Fig. 1Dc,d). We considered that the absence of TD-RPE ramping in the case with SR could provide an explanation of the observed gradual fading of DA ramping over daily training of reward navigation [7]. Indeed, we found that when representation of states towards reward was initially punctate but gradually became closer to SR through learning of SR features, TD-RPE ramping appeared in initial/early trails but then gradually faded away (Fig. 1Fc, compare to Fig. 4a of [7]). Notably, TD-RPE prominently ramped at a later (reward-proximal) phase in initial trials (Fig. 1Fb) but more constantly ramped in subsequent trials (Fig. 1Fa), resembling the observed patterns of DA ramping (Fig. 1j and 1e of [7]).

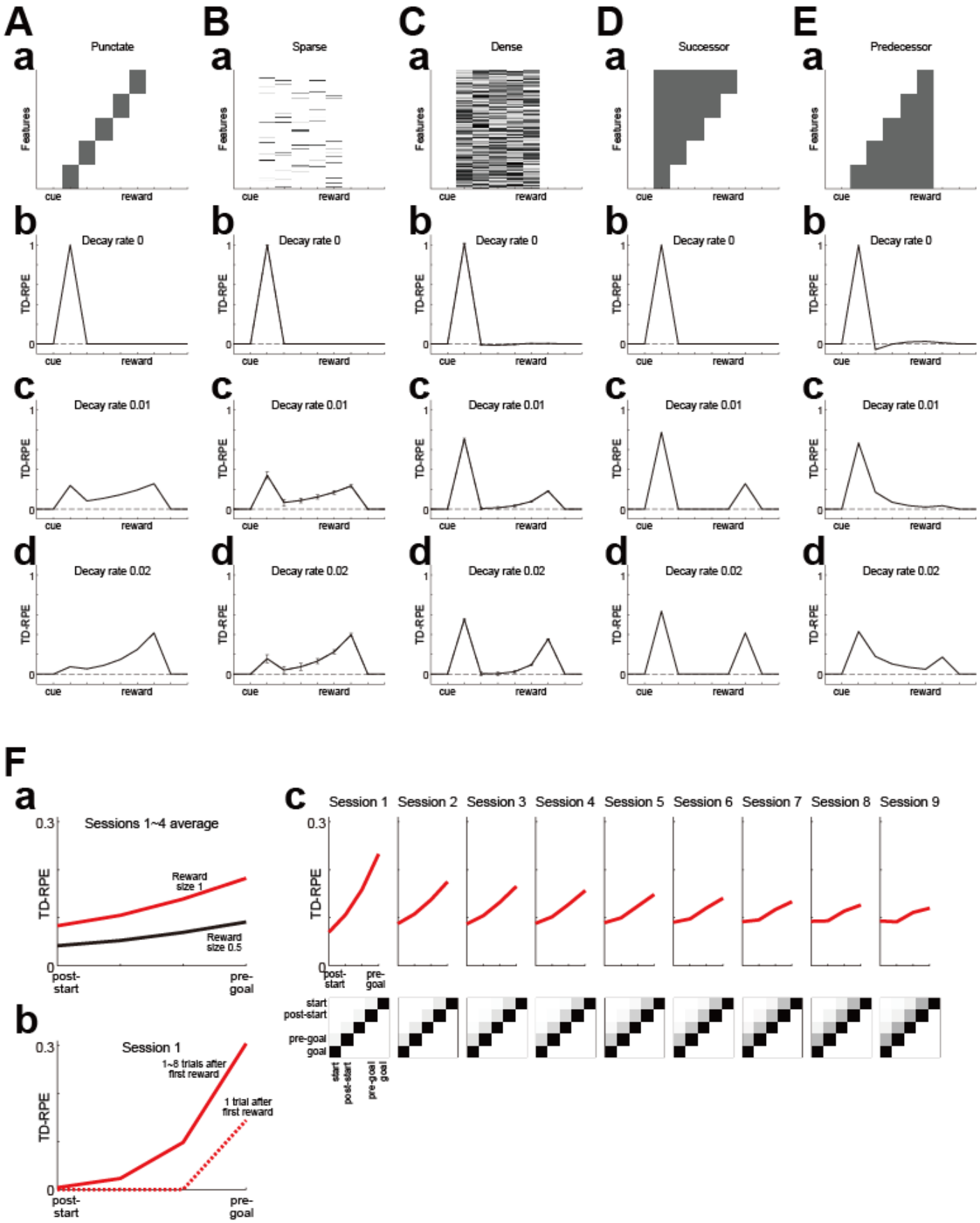
A
B
C
D
E
a
Punctate
Sparse
Dense
Successor
Predecessor
Features
cue
reward
b
Decay rate 0
c
Decay rate 0.01
d
Decay rate 0.02
TD-RPE
F
Sessions 1~4 average
Reward size 1
Reward size 0.5
post-start
pre-goal
Session 1
1~8 trials after first reward
1 trial after first reward
Session 2
Session 3
Session 4
Session 5
Session 6
Session 7
Session 8
Session 9
start
post-start
pre-goal
goal

Figure 1

Figure 1
Effects of value-decay on the TD-RPE pattern under various state representations. **(A-E)** TD-RPE generated in a simulated cue-reward association learning task after learning became almost steady in cases with various state representations and rates of value(-weight)-decay. Time discount factor was set to 1 in all the cases. (A) Case where five timings from cue to reward were separately represented in a punctate manner, as schematically indicated in (a). The rate of value-decay was set to 0 (no decay) (b), 0.01 (c), or 0.02 (d). (B,C) Cases where the five timings were represented by sparse (B) or dense (C) pseudo-random feature vectors (indicated in (a)). The lines and error-bars in (b-d) indicate the mean±SD across 100 simulations. (D,E) Case where the five timings were represented by the successor representation (SR) (D) or the predecessor representation (PR) (E). **(F)** TD-RPE ramping in simulated reward navigation gradually faded away when SR was gradually shaped through TD learning of SR features. **(a)** TD-RPE from post-start to pre-goal states averaged over 1st ~ 4th sessions, with each session consisting of 40 trials. The red and black lines indicate cases with reward size 1 and 0.5, respectively. Compare to DA ramps in navigation towards big and small rewards in Fig. 1e of [7]. **(b)** TD-RPE averaged over 1st ~ 8th trials (red solid line) or at the 1st trial (red dotted line) after the initial trial in which reward (size 1) was obtained. Compare to DA ramps in early trials after first reward was obtained in Fig. 1j of [7]. **(c)** TD-RPE (top) and state representation (feature) matrix (bottom) in 1st ~ 9th sessions. Compare with the experimentally observed fading of DA ramps over daily training in Figure 4a of [7].

As so far shown, value(-weight)-decay generally caused sustained TD-RPE at reward after learning completion, and also ramping TD-RPE towards reward except for the case with SR, but their degrees heavily depended on state representation. Given that dense representation, in which feature vectors were much overlapped, caused small ramping and SR, in which feature vectors are nested and so in a sense completely overlapped, caused no ramping, it could be guessed that the degrees of ramping depend, negatively, on the degree of overlaps between feature vectors. To test this, we examined a case with 'predecessor' representation (PR) (c.f., [63]), where the states were represented by past, rather than future, occupancies of other (and self) states. In this case, value-weight-decay caused rather small sustained TD-RPE at reward and little ramping-up but instead ramping-down (Fig. 1E). Because the PR feature vectors had the same degree of overlaps as the SR feature vectors had, this result suggests that not only the degree of overlaps between feature vectors but also their temporal order affects the pattern of TD-RPE.

Hierarchical RL model consisting of circuits with and without value-decay

As mentioned in the Introduction, it was shown that CDh neurons rapidly learned the cue value within a session but forgot it in subsequent days while CDt neurons gradually developed stable value-memory [43]. It was also shown that DA neurons in rmSNc and clSNc, which project to CDh and CDt, respectively, forgot and retained predicted-value-memory [44]. Specifically, rmSNc DA neurons developed TD-RPE response during the cue-reward association task but had lost it at the subsequent passive viewing task, whereas clSNc DA neurons retained TD-RPE response in the latter task. These results appear to be potentially explained if CDh-rmSNc circuit implements RL with a large learning rate and value-decay whereas CDt-clSNc circuit implements RL with a small learning rate and no value-decay. But how do these two circuit interact? It was found [44] that the memory-retain type clSNc DA neurons did not respond to unpredicted reward. The authors

suggested that these DA neurons may instead receive 'conditioned reinforcement' information from the forget/update type rmSNc DA neurons, which responded to unpredicted reward, potentially through the superior colliculus [44]. Such conditioned reinforcement information might also be sent through the suggested 'spiral' structure of the striatum-midbrain circuits [64,65].

Based on these findings and considerations, we constructed a hierarchical RL model (c.f., [52-57]) consisting of coupled two circuits, the circuit M, modeling CDh-rmSNc, and the circuit L, modeling CDt-clSNc (Fig. 2A). The circuit M had a (relatively) large learning rate and value-decay whereas the circuit L had a small learning rate and no value-decay. Also, the circuit M receives external reward information, whereas circuit L does not receive direct reward input but receives the information of upcoming state value from the circuit M. We simulated the cue-reward association learning task by this model, assuming the punctate state representation. As learning progressed, TD-RPEs in both circuits M and L responded to cue while TD-RPE in the circuit M additionally showed sustained response to reward (Fig. 2Ba). These patterns look similar to the reported DA patterns in clSNc and rmSNc (Figure 3A-E of [44]) including a small apparent response of rmSNc DA to reward although it could just reflect contributions of initial trials.

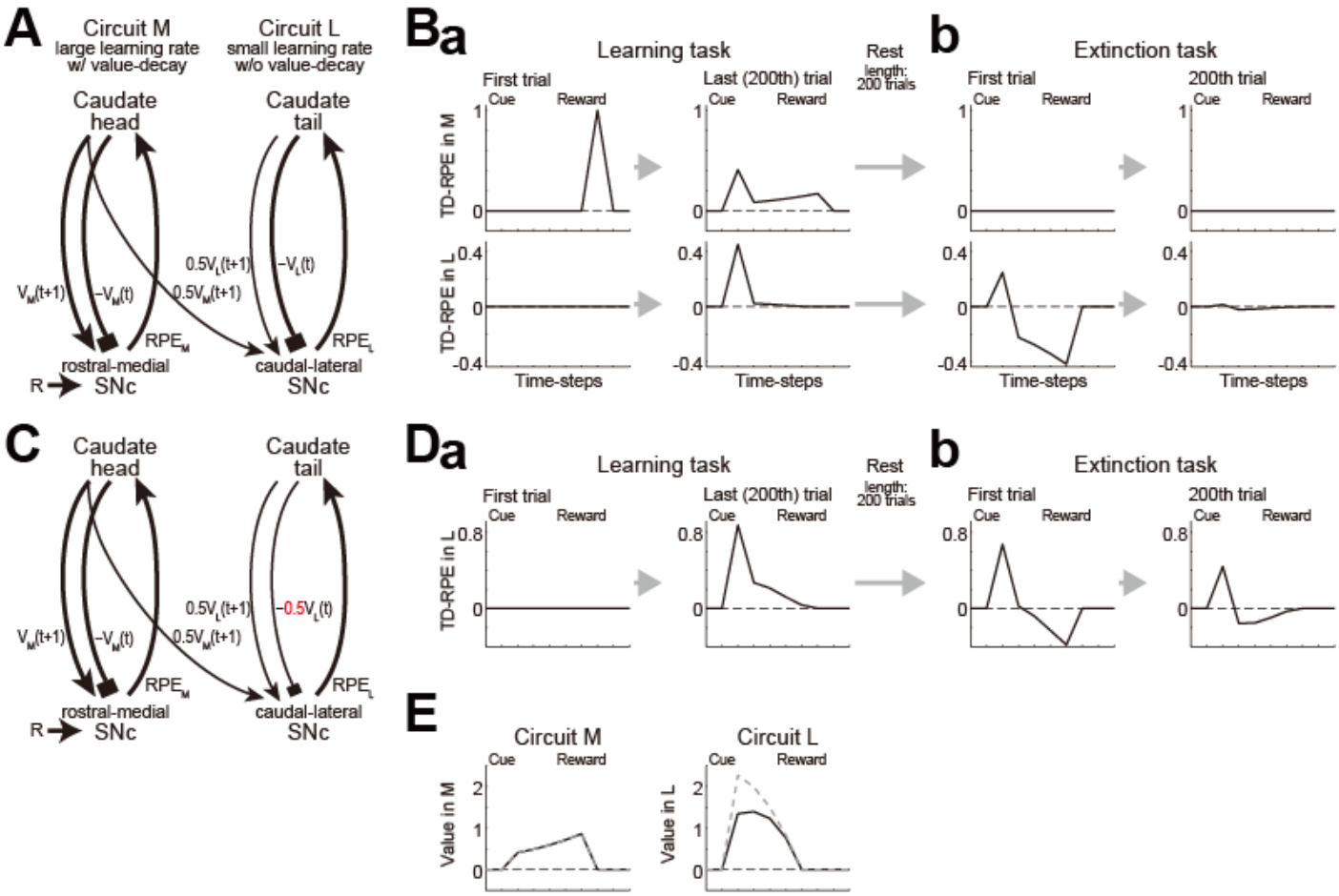

Figure 2

Figure 2
Hierarchical RL model of striatum-midbrain circuits with differences in learning rate and value-decay. **(A)** Hierarchical RL model. The circuit M, with a large learning rate and value-decay, and the circuit L, with a small learning rate and no value-decay, modeled the CDh-rmSNc and CDt-clSNc circuits, respectively. **(B)** TD-RPE generated in the model's circuit M (top) and circuit L (bottom) during a simulated learning task (a) and a subsequent extinction task (b). Compare with the experimental results of DA neuronal activities in Figure 3A-E of [44]. **(C)** The revised model with the strength of the negative "current state value" input in the circuit L reduced (halved: $-0.5V_L(t)$), as indicated by the red color, so that it matched the strength of the "upcoming state value" input in the circuit L ($0.5V_L(t+1)$) only, corresponding to the case where effective time discount factor (i.e., upcoming-value/current-value ratio) was 1 in the absence of the conditioned M→L input. **(D)** TD-RPE generated in the circuit L of the revised model. **(E)** State values in the circuit M (left) and circuit L (right) of the revised model at the last (200th) trial of the learning task (black solid lines). The gray dashed lines indicate the cases where the number of trials was increased to 400 (overlapped with the black solid line in the left panel). Compare with the experimental results of caudate neural activities in Figure 3B,C of [43].

In the experiment [44], after the learning task, monkeys engaged in free viewing and passive viewing sessions, in which multiple stimuli were simultaneously or sequentially presented without contingent reward, and memory retention was examined. In our model, after the learning task (200 trials), we simulated a rest period with the same duration (corresponding to 200 trials), and then an extinction task (200 trials), in which the learned cue was presented without reward. At the beginning of the extinction task, TD-RPE in the circuit M disappeared whereas TD-RPE upon cue in the circuit L was retained (Fig. 2Bb-left), in line with the disappearance and retention of rmSNc and clSNc neuronal response, respectively [44]. However, the positive cue response of TD-RPE in the circuit L was followed by a large negative response (Fig. 2Bb-left-bottom) and these responses almost disappeared after extinction learning (Fig. 2Bb-right-bottom), deviating from the experimental result.

We explored how this discrepancy could be mitigated, and found that reducing the coefficient of the negative "current state value" input ($-V_L(S_i)$) in the circuit L (Fig. 2C: where the coefficient was halved (from 1 to 0.5)) was effective, as shown in Figure 2D, although some negative response still remained. In the original model (Fig. 2A), this coefficient was set to match the sum of coefficients of the "upcoming state value" inputs from the circuit M ($0.5V_M(S_{i+1})$) and the circuit L itself ($0.5V_L(S_{i+1})$) (i.e., 0.5 + 0.5 = 1), corresponding to the case where effective time discount factor (i.e., upcoming-value/current-value ratio) was 1 in the presence of the conditioned M→L input. The revised model (Fig. 2C) instead corresponded to the case where effective time discount factor was 1 in the absence of the M→L input. The better reproduction of the experimentally observed long-term retention of clSNc DA response by the latter model indicates that the actual striatum-DA circuits are tuned in that way: in the extinction task, in the circuit L, the magnitude of the negative "current value" input is not much larger than the magnitude of the positive "upcoming value" input so that TD-RPE does not become quite negative, ensuring long-term retention of response.

If that is the case, in the presence of the M→L input, effective time discount factor exceeded 1 in the circuit L, and so delay should cause amplification, rather than discounting, of reward value. Indeed, the

value in the circuit L at the end of the learning task showed such a pattern, i.e., backwardly ramped towards the early phase of object presentation (Fig. 2E-right, black solid line). By contrast, the value in the circuit M showed a ramp towards the late phase, reflecting temporal discounting caused by value-decay (Fig. 2E-left, black solid line). When the number of trials of the learning task was increased, the cue response in the circuit L was further increased while the activity of the circuit M remained unchanged (Fig. 2E, gray dashed lines). Crucially, these temporally increasing and decreasing values in the circuits M and L appear to match the observed patterns of CDh and CDt neuronal activities, respectively (Figure 3B,C of [43]), providing a possible mechanistic explanation. Functionally, such a value amplification in the circuit L is considered to contribute to long-term value retention after reward extinction.

Distributional RL algorithms with and without value-decay

Recent studies suggested that the cortico-basal ganglia-DA system implements distributional RL [27-31,33]. Given that D1-direct-pathway and D2-indirect-pathway striatal projection neurons (SPNs) predominantly learn from positive and negative TD-RPE, respectively (c.f. [27,58,59,66]), they can develop positively biased D1 value and negatively biased D2 value [27] and thus collectively have information of reward distribution. There are two (extreme) possibilities on how TD-RPE is calculated from D1 and D2 values. The first one is that D1 and D2 values are combined (summed) into an integrated value, from which an integrated TD-RPE is calculated and used for updates of both D1 and D2 values. We refer to this as the integRPE algorithm (Fig. 3A left), which is a TD version (extension) of the Actor learning Uncertainty (AU) model [27]. The second possibility is that pathway-specific TD-RPEs are calculated based on either D1 value or D2 value only, and each pathway is updated using such a segregated TD-RPE. We refer to this as the segreRPE algorithm (Fig. 3A right), which is a TD version (extension), with simplifications, of the reflected EDRL (REDRL) model [29] (an extension of the Expectile Distributional RL (EDRL) [67]).

We simulated a probabilistic cue-reward association learning task, in which reward of either size 1 or 3 was obtained with equal probabilities, using each of these two algorithms. When the integRPE algorithm was used, D1 values became positive whereas D2 values became mostly negative (Fig. 3Ba-leftmost), and the magnitudes of both of these values grew unboundedly (compare the values at 2000th trial with those at 200th trial in Fig. 3Ba-left), while the magnitudes of the integrated value and TD-RPE did not. Such unbounded value growth, which seems biologically implausible, was prevented if value-decay was incorporated (Fig. 3Ba-right). In this sense, value-decay is necessary for the integRPE algorithm, and indeed value-decay was assumed in the original AU model [27]. In contrast, when the segreRPE algorithm was used, both D1 and D2 values became positive, and they did not grow unboundedly even without value-decay (Fig. 3Bb). Thus, value-decay is not necessary for the segreRPE algorithm, and it was not assumed in the original REDRL model [29].

We also simulated a deterministic cue-reward association task, in which reward of size 2 (equal to the mean of the reward sizes in the abovementioned probabilistic task) was always obtained, again using each of the two algorithms. When the integRPE algorithm (with value-decay) was used, D1 and D2 values became both positive but separated; D1 value became larger than D2 value (Fig. 3Ca). In contrast, when the segreRPE algorithm (without value-decay) was used, D1 and D2 values converged to almost the same positive value

(Fig. 3Cb-left). This way, the segreRPE algorithm better discriminated the probabilistic reward and the deterministic reward than the integRPE algorithm (although the integRPE algorithm could still encode the reward split (c.f. [27]) as apparent from the result for a task with a larger reward split (0 or 4) (Fig. 3Db)). Notably, if value-decay was assumed in the segreRPE algorithm, D1 and D2 values no longer converged to an almost same value but to different values (Fig. 3Cb-right). Thus, value-decay is not only unnecessary but harmful for the segreRPE algorithm to achieve the ideal differential coding of the different reward distributions.

The integRPE algorithm with value-decay generated ramping TD-RPE (Fig. 3Ea) whereas the segreRPE algorithm without value-decay generated only abrupt TD-RPE (Fig. 3Eb). Also, in the integRPE algorithm with value-decay, D1 value ramped up and D2 values also (less prominently) ramped up except for the last part, whereas D1 and D2 values are largely flat in the segreRPE algorithm without value-decay (Fig. 3B-D). Regarding the coding property, the segreRPE algorithm without value-decay better discriminated the different reward distributions, and so could realize better distribution-coding than the integRPE algorithm with value-decay. Conversely, the integRPE algorithm with value-decay encoded different distributions with the same mean more similarly, and so could realize better mean-coding than the segreRPE algorithm without value-decay. We quantified the strength of distribution- and mean-coding of the two algorithms (c.f., [29]), through simulations assuming some noise (see the Methods for details). The results confirmed and refined the abovementioned conjectures. As for the strength of distribution-coding (Fig. 3F), the score of the segreRPE algorithm without value-decay was high for all the time-steps from cue to reward whereas the score of the integRPE algorithm with value-decay was low except for the last time-step, regardless of the noise levels. Regarding the strength of reward mean-coding (Fig. 3G), while the score of the segreRPE algorithm without value-decay was nearly constant across time-steps, the score of the integRPE algorithm with value-decay showed a gradual increase, surpassing the segreRPE when the noise was small (Fig. 3G-left), and a drop at the last time-step. The gradual increase in the integRPE's mean coding reflects the increase (ramp) of the D1/D2 values and a resulting improvement in the signal-to-noise ratio.

Given these differential properties, here we propose that the integRPE and segreRPE algorithms can coherently explain recent findings on DA patterns and coding properties in the nucleus accumbens shell. As for the DA pattern, recent work [6] found that DA in lateral nucleus accumbens shell (lAcbSh) showed an abrupt cue response whereas DA in medial accumbens shell (mAcbSh) showed a sustained/ramping pattern. These DA patterns appear to be consistent with a previous finding [5] that medial VTA DA neurons exhibited more positive ramps than lateral VTA DA neurons. Regarding the coding property, another recent work [29] showed that distributional coding was strongest in lAcbSh and weaker in mAcbSh (although $n = 1$ for mAcbSh). As for mean coding, the strength in lAcbSh was relatively constant between cue and pre-reward timings whereas the strength in mAcbSh ramped and surpassed that of lAcbSh [29]. We argue that these results for DA patterns and AcbSh coding properties can be coherently explained if lAcbSh and mAcbSh implement the segreRPE algorithm without value-decay and the integRPE algorithm with value-decay, respectively (Fig. 3H). Crucially, this explanation can reconcile the controversy on whether ramping DA in mAcbSh/medial VTA represents value [6] or TD-RPE [5]. Specifically, because ramping TD-RPEs in the case with value-decay can be proportional to state values, except for cue response, according to the formulae that we previously derived [9], the ramping DA in mAcbSh could in fact represent both TD-RPE and value.

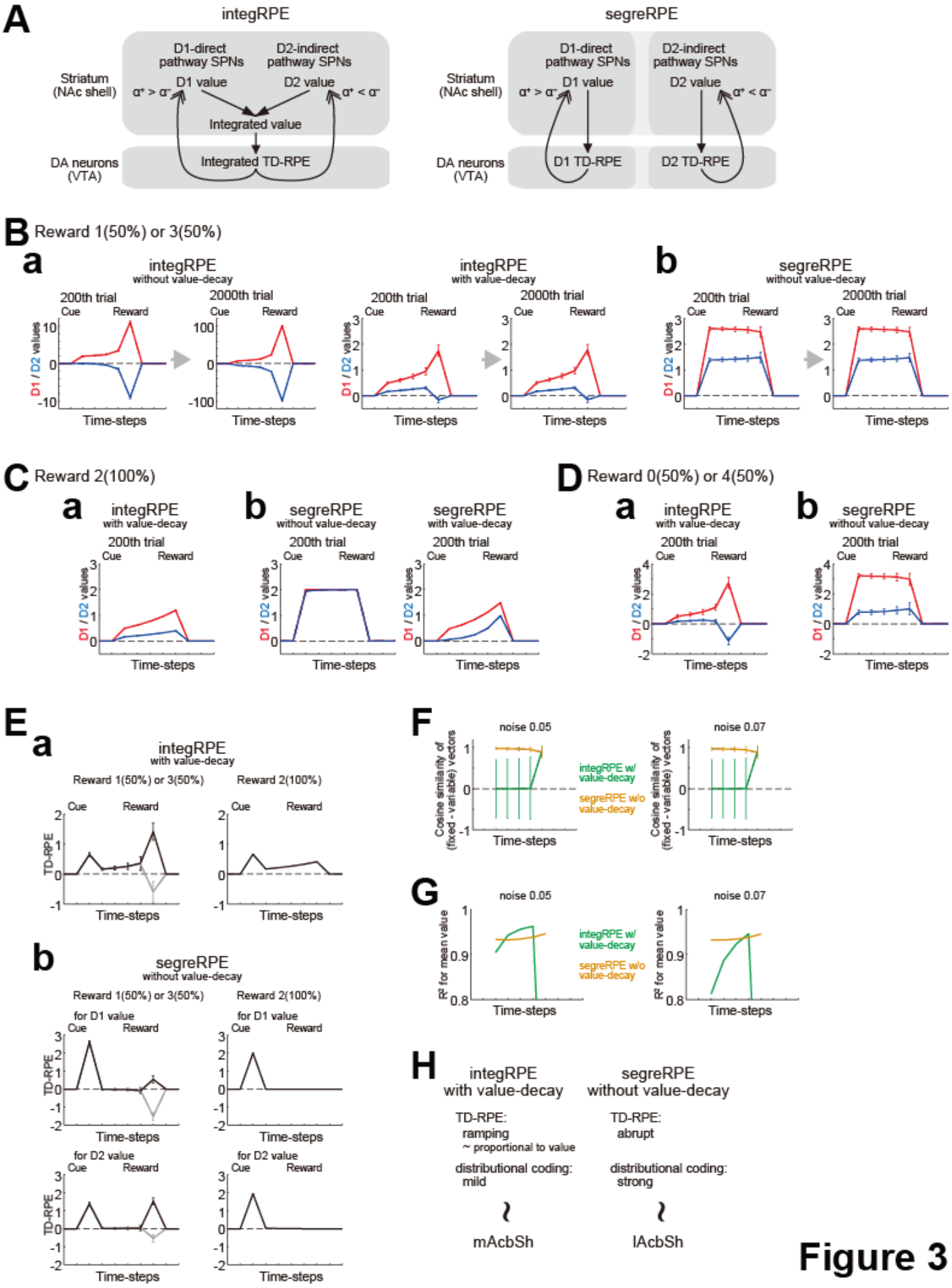
A
integRPE
D1-direct pathway SPNs
D2-indirect pathway SPNs
Striatum (NAc shell)
D1 value
D2 value
α⁺ > α⁻
α⁺ < α⁻
Integrated value
DA neurons (VTA)
Integrated TD-RPE
segreRPE
D1-direct pathway SPNs
D2-indirect pathway SPNs
Striatum (NAc shell)
D1 value
D2 value
α⁺ > α⁻
α⁺ < α⁻
DA neurons (VTA)
D1 TD-RPE
D2 TD-RPE
B Reward 1(50%) or 3(50%)
a
integRPE without value-decay
integRPE with value-decay
b
segreRPE without value-decay
200th trial
2000th trial
Cue
Reward
D1 / D2 values
Time-steps
C Reward 2(100%)
a
integRPE with value-decay
b
segreRPE without value-decay
segreRPE with value-decay
200th trial
D Reward 0(50%) or 4(50%)
a
integRPE with value-decay
b
segreRPE without value-decay
E
a
integRPE with value-decay
Reward 1(50%) or 3(50%)
Reward 2(100%)
TD-RPE
Time-steps
b
segreRPE without value-decay
for D1 value
for D2 value
F
noise 0.05
noise 0.07
Cosine similarity of (fixed - variable) vectors
integRPE w/ value-decay
segreRPE w/o value-decay
Time-steps
G
noise 0.05
noise 0.07
R² for mean value
integRPE w/ value-decay
segreRPE w/o value-decay
Time-steps
H
integRPE with value-decay
segreRPE without value-decay
TD-RPE: ramping
~ proportional to value
TD-RPE: abrupt
distributional coding: mild
distributional coding: strong
mAcbSh
lAcbSh

Figure 3

Figure 3

Behavior of different types of distributional RL algorithms with and without value-decay in probabilistic or deterministic cue-reward association learning tasks. **(A)** Different types of distributional RL algorithms. *Left*: integRPE, where D1 and D2 values are combined into the integrated value, from which the integrated TD-RPE is calculated and used for updates of both D1 and D2 values. *Right*: segreRPE, where D1 [/D2] TD-RPE is calculated only from D1 [/D2] values, and used only for updating D1 [/D2] value. **(B)** D1 values (red lines) and D2 values (blue lines) in a probabilistic task where reward was 1 or 3 with equal probabilities. The lines and error-bars indicate the mean±SD across 100 simulations (same applied to panels in (C,D)). **(a)** Results for the integRPE algorithm. Cases without value-decay (left two panels) and with value-decay (right two panels) at 200th and 2000th trials are shown. **(b)** Results for the segreRPE algorithm. Value-decay was not assumed, and cases at 200th and 2000th trials are shown. **(C)** D1 and D2 values in a deterministic task (reward: 2) in the integRPE algorithm with value-decay (a) and the segreRPE algorithm without value-decay (b-left: red and blue lines are almost overlapped) or with value-decay (b-right). **(D)** D1 and D2 values in a probabilistic task (reward: 0 or 4 with equal probabilities) in the integRPE algorithm with value-decay (a) and the segreRPE algorithm without value-decay (b). **(E)** TD-RPE in the probabilistic task (reward: 1 (gray lines) or 3 (black lines)) or in the deterministic task. The lines and error-bars indicate the mean±SD across simulations corresponding to each condition (reward size) within total 100 simulations. **(a)** Results for the integRPE algorithm with value-decay. **(b)** Results for the segreRPE algorithm without value-decay. Both TD-RPE for D1 values and TD-RPE for D2 values are shown. **(F)** Strength of reward distribution-coding quantified by an average cosine similarity of the difference in [D1 value, D2 value] vectors, at each time-step, between the condition where reward was always 2 ("fixed") and the condition where reward was 1 or 3 with equal probabilities ("variable"). The green and orange lines indicate the integRPE algorithm with value-decay and the segreRPE algorithm without value-decay, respectively, and the lines and error-bars indicate the mean±SD across all possible pairs from 1000 simulations. **(G)** Strength of reward mean-coding quantified by R^2 of correlation between the reward mean and the D1 value and D2 value, normalized across three conditions where reward was always 0, always 2, or 1 or 3 with equal probabilities. Note that the vertical axis starts at 0.8, and also that the result for the integRPE at the last (fifth) time-step was outside the drawing range in both panels. **(H)** Proposed correspondence between the types of distributional RL algorithms, the patterns of TD-RPE, the strengths of distributional coding, and the regions of AcbSh.

Discussion

We have shown how heterogeneous DA patterns and various RL computations can be linked by value-decay. Specifically, we have shown that (i) value-decay causes ramping TD-RPE under non-predictive representations but not SR, explaining the observed gradual fading of DA ramping over experience, (ii) the hierarchical RL model consisting of circuits with and without value-decay explains regional differences in DA and striatal activity patterns and memory flexibility/stability, and (iii) the distributional RL algorithms with and without value-decay coherently explain regional differences in DA patterns and strengths of distributional coding. Below we discuss each of these, with limitations and perspectives. Our results suggest that there exist within-striatum differences, or more specifically, medial-lateral gradients in the degree of value-decay, which presumably originate from heterogeneities in forgetting-related molecules and regulate regionally flavored RL computations.

Dependence of ramping and effective temporal discounting on value-decay and state representation

Our results suggested that gradual fading of DA ramping as training progressed in reward navigation [7] could result from gradual formation of SR. Persistence of DA ramping in another wheel-running task without spatial cues in the same study [7] could also be explained by no formation of SR due to the lack of spatial cues. Since SR is more automatic than explicit model-based planning, this can be in line with the authors' argument that ramping DA appears when animals use a cognitive map, although what causes ramping in our model is model-free RL under punctate or sparse representation, which could operate in parallel with model-based planning. The authors suggested that the ramping DA was not RPE because it quickly appeared in the next trial after the first reward was obtained. But quick RPE-like DA response was reported previously [44], and the observation that DA ramping appeared in a late timing in initial trials but appeared from the beginning in later trials (Fig. 1j and 1e of [7]) matches the property of TD-RPE in our simulations (Fig. 1Fa,b). A limitation of our present model/simulation is that passive state transition and active transition by voluntary action [68], as well as state-value/critic learning and action-value/actor learning [21,69-72], are not distinguished. Incorporation of them would be an important future direction.

Value-decay caused a decrease of TD-RPE upon cue. This can be regarded as effective temporal discounting of cue value, even though genuine temporal discounting (time discount factor <1) was not assumed [11]. Our present simulation results suggest that the degree of effective temporal discounting by value-decay depends on state representation (Fig. 4). Given these, caution would be needed when estimating the time discount factor from DA patterns [73,74].

Figure 4
Dependence of effective temporal discounting on value-decay and state representation. The horizontal axis indicates the time-step of cue relative to reward (0), and the vertical axis indicates the cue value, i.e., TD-RPE upon cue. Decay rate was 0.01 (A) or 0.02 (B). The gray dotted, gray solid, black solid, black dotted, and black dashed lines indicate the cases with the punctate, sparse, dense, successor (SR), and predecessor (PR) representations, respectively. In the cases of sparse and dense representations, the lines and error-bars indicate the mean±SD across 100 simulations.

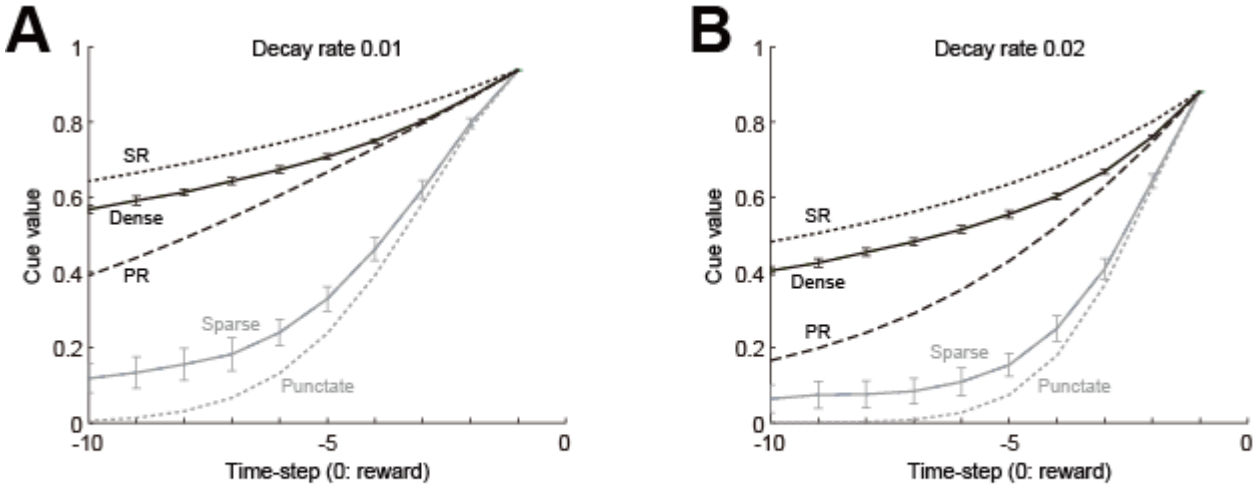

Figure 4

Hierarchical RL circuits with differences in learning rate and value-decay

We considered hierarchical coupling of the upstream circuit with a large learning rate and value-decay and the downstream circuit with a small learning rate and no value-decay to explain the flexible and stable value memories in CDh-rmSNc and CDt-clSNc, respectively. In order to reproduce the observed stable sustain of value-memory in CDt-clSNc [43], further assumption was needed. Specifically, the negative "current state value" input in the downstream circuit needed to be not strong, balancing only the "upcoming state value" input within the circuit, omitting the "upcoming state value" input from the upstream circuit that represents the conditioned reinforcement. With this assumption, during the learning task in which the conditioned reinforcement input existed, the effective time discount factor (upcoming-value/current-value ratio) of the downstream circuit exceeded 1, resulting in amplification, rather than discounting, of cue value. The amplified cue value in the model indeed matched the observed rather high cue response of CDt neurons. Such a value amplification may enable a formation of strong habit for efficient behavior, but could also make it difficult to change/discard habits, potentially causing a risk of addiction.

In the experiments, caudate and DA neurons showed response not only to stimuli associated with reward but also, to lesser extents in most cases, to those associated with no reward. Such responses are not explained by our present model. They could potentially relate to generalization or context [75-77], and a possible future direction is to consider models incorporating feature-based representation [78] that enables generalization. Response to no-reward stimuli in CDt-clSNc could also reflect saliency. Specifically, primate CDh and CDt have homologies to rodent DMS [79] and tail of striatum (TS) [80-82], respectively, although CDt has stronger vision-selectivity than TS [80,81]. and apparently dissimilar stable-value-encoding in CDt [43,44] and potential-threat-encoding in TS [83-85] could be commonly understood as encoding of learned stimulus saliency [82].

Along with CDt/TS, rodent dorsolateral striatum (DLS) has also been suggested to be crucial for stimulus-response behavior [86]. Since DA in DLS responds to unpredicted reward, different from TS, DLS could be considered as an intermediate between DMS and TS. Then, our model predicts smaller learning rate and value-decay in DLS than in DMS. In cue-reward association task with multiple reward sizes [87], DA in

DMS showed positive and negative response to large and small reward, respectively, as usually expected for RPE, but DA in DLS showed positive response to both: even to small reward. This could potentially be explained by a small learning rate: RPE for small reward can be positive if learning is still incomplete, as confirmed in our simulations (Fig. 5).

A recent study [74] suggested that DA dynamics is fast in DLS, slower in DMS, and even slower in the ventral striatum (VS). At first glance, this looks contradictory to our proposal of smaller learning rate and value-decay in DLS than in DMS. However, these two might be the opposite sides of the same coin. DLS may steadily learn the value of individual short action-elements (movements), and DA/RPE there is abrupt since there is no value-decay (value decreases due to negative RPE). In contrast, DMS (and VS) may quickly learn and forget the value of macro-level actions, which reflect the context, such as the current reward richness/poorness, coming from the environmental model and can thus fluctuate in a slow time scale.

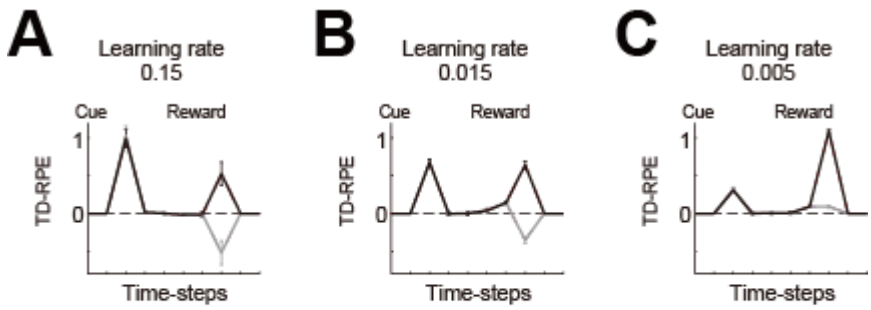

Figure 5

Figure 5
TD-RPE in a probabilistic cue-reward association task (reward: 0.5 (gray lines) or 1.5 (black lines) with equal probabilities) at 100th trial in the cases with different learning rates: 0.15 (A: learning is almost completed), 0.015 (B: learning is still incomplete), or 0.005 (C: learning is still incomplete). The lines and error-bars indicate the mean±SD across simulations corresponding to each condition (reward size) within total 100 simulations. Value-decay was not assumed, and the dense state representation (Fig. 1Ca) was assumed. Notably, if the states are instead represented in the punctate or sparse manner, in the period in which learning is incomplete, TD-RPE appears prominently at intermediate states between cue and reward and hardly or only limitedly at cue, deviating from the observed DA pattern in DLS [87].

Different types of distributional RL with and without value-decay

We have presented two algorithms of distributional RL, integRPE and segreRPE, summarizing and extending the previous models [27-29], and shown that their implementations in different striatal regions, mAcbSh and lAcbSh, respectively, can coherently explain the regional differences in DA patterns [6] and the strengths of distributional coding [29]. Note, however, that the result for coding strength is currently based on limited sample (n = 1 for mAcbSh) [29], and so it could rather be regarded as a prediction that awaits further experiment.

Are there any merits of having both algorithms? The merit of the segreRPE algorithm (REDRL model [29]) is clear: it achieves better discrimination of deterministic and probabilistic rewards, and more generally, the REDRL model enables complete learning/representation of the reward distribution [29]. Value-decay is not allowed for ideal performance (Fig. 3Cb). What can the integRPE algorithm (or its non-TD original, AU model [27]) be beneficial for? One possibility is an anatomical merit, or rather a constraint. The segreRPE algorithm requires that segregated RPEs for D1 and D2 values are calculated and utilized without intermixing, but it may be difficult because of the widespread dense axonal projections of DA neurons [88] and intermingled placements of D1 and D2 SPNs in most striatal regions albeit with exceptions [89-92]. The integRPE algorithm can be more simply implemented through broadcasting DA signals.

The integRPE algorithm might also have functional merits. Because individual DA neurons encode the integrated RPE, its information could be easily used. Moreover, the integRPE algorithm requires value-decay for stability (Fig. 3Ba), and then, as discussed in the Results, value-decay could achieve simultaneous representation of the integrated TD-RPEs and the integrated state values. In situations where animals/humans need to quickly learn from a small number of samples and/or the environment is rapidly changing, accurate representation of the entire reward representation may not be possible even by the segreRPE algorithm. Then, the simply implementable integRPE algorithm may suffice, as it can still learn reward distribution to a certain degree [27]. The integRPE algorithm may thus be compatible with large learning rate and value-decay, as we suggested for CDh and its homolog DMS. This is potentially in line with the weaker distributional coding in DMS than in lAcbSh [29].

Distributional coding was shown to be relatively weak also in DLS [29]. This apparently seems at odds with our suggestion of no value-decay in DLS, which may be compatible with the segreRPE algorithm. However, the weak distributional coding in DLS can be due to incomplete learning because of small learning rate, as discussed above. Alternatively, it can be due to unbalanced D1 and D2 values/RPEs, more specifically, dominance of D2 over D1 values/RPEs, which was suggested for explaining the positive shift of RPE in DLS[87] and is potentially in line with the stronger projection from the primary motor cortex to D2 than D1 pathway [93,94].

Methods

Effects of value-decay under various state representations

We simulated the behavior of RL agent in a cue-reward association learning task. In each trial, upon seeing the cue stimulus, the agent entered state S_1, and then sequentially and deterministically transitioned to state S_2, ..., and S_n at each time step. At state S_n, the agent received a reward of a fixed size ($R_n = 1$) deterministically (for Figures 1 and 4) or of different sizes ($R_n = 0.5$ or 1.5) with equal probabilities (for Figure 5), and at the other states, the agent did not receive any reward ($R_i = 0$ ($i \neq n$)). The number of states (n) was set to 5 in most simulations except for those for Figure 4, in which n was varied from 1 to 10. The number of trials was set to 200 for Figures 1 and 4, and 100 for Figure 5.

We considered an agent having a learning system, which had the predicted value of each state, $V(S_i)$ ($i = 1, ..., n$), initialized to all 0. At each time-step/state, TD-RPE:

$$\delta_i = R_i + \gamma V(S_{i+1}) - V(S_i) \quad \text{(if } i = n\text{, the } \gamma V(S_{i+1}) \text{ term was dropped)}$$

was calculated, where γ was the time discount factor and was set to 1 in all the simulations. $V(S_i)$ was then updated as

$$V(S_i) \leftarrow V(S_i) + a\delta_i,$$

where a was the learning rate and was set to 0.15. At each time-step/state, the predicted values of all the states decayed [9] as

$$V(S_i) \leftarrow (1 - \kappa)V(S_i) \quad (i = 1, ..., n),$$

where κ was the decay rate and was set to 0, 0.01, or 0.02. Note that value-decay was imposed at every time-step only during the task trials.

We next considered cases where each state, S_i ($i = 1, ..., n$), was represented by a 100 dimensional feature vector:

$$\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,100}).$$

We examined cases with two different types of representation: sparse representation and dense representation. In the sparse representation, for each feature vector $\boldsymbol{x}_i$, only a subset of elements (10 out of 100), which were pseudo-randomly selected, were drawn from [0 1] uniform pseudo-random numbers while the remaining elements were 0, and subsequently each feature vector was scaled so that its norm became 1. In the dense representation, all the elements of each feature vector $\boldsymbol{x}_i$ were drawn from [0 1] uniform pseudo-random numbers, and subsequently each feature vector was scaled so that its norm became 1. We also considered cases where the states were represented by the successor representation (SR) [48,49] or the predecessor representation (PR) (c.f., [63]). In these case, the states were represented by future or past occupancies of other (and self) states (in $n = 5$ dimension):

$$x_{i,j} = \gamma^{j-i} \text{ (for } j \geq i\text{) or 0 (for } j < i\text{) in SR, or}$$

$$x_{i,j} = \gamma^{i-j} \text{ (for } j \leq i\text{) or 0 (for } j > i\text{) in PR,}$$

where γ was set to 1. In all these cases (sparse, dense, SR, and PR), predicted value of the state (i.e., state value) was calculated as an inner product of the features and a common weight vector

$$\boldsymbol{w} = (w_1, ..., w_m) \quad (m = 100 \text{ for the sparse and dense representations and } m = 5 \text{ for the SR and PR})$$

with subtraction of constants:

$V(S_i) = \boldsymbol{w}\cdot\boldsymbol{x}_i - \boldsymbol{w_0}\cdot\boldsymbol{x}_i = w_1x_{i,1} + ... + w_mx_{i,m} - w_0(x_{i,1} + ... + x_{i,m})$,

where $\boldsymbol{w_0} = (w_0, ..., w_0)$ was a constant vector, to which $\boldsymbol{w}$ was initialized, and w_0 was set to 0.5. At each time-step/state, TD-RPE:

$\delta_i = R_i + \gamma V(S_{i+1}) - V(S_i)$ (if $i = n$, the $\gamma V(S_{i+1})$ term was dropped)

was calculated, where γ was set to 1 in all the simulations. The weight vector $\boldsymbol{w}$ was updated as

$\boldsymbol{w} \leftarrow \max(0, \boldsymbol{w} + a_+\delta_i\boldsymbol{x}_i/||\boldsymbol{x}_i||^2)$ if $\delta_i \geq 0$ or

$\boldsymbol{w} \leftarrow \max(0, \boldsymbol{w} + a_-\delta_i\boldsymbol{x}_i/||\boldsymbol{x}_i||^2)$ if $\delta_i < 0$,

where the max operation ensured that the weights remained non-negative, and a_+ and a_- were the learning rates for positive and negative TD-RPEs, respectively, which were set to 0.15 and 0.075 in Figures 1 and 4 and both 0.15, 0.015, or 0.005 in Figure 5. At each time-step/state, all the weights decayed to their initial values:

$w_i \leftarrow w_0 + (1 - \kappa)(w_i - w_0)$ $(i = 1, ..., m)$,

where κ was set to 0, 0.01, or 0.02 (note that value-decay was imposed at every time-step only during the task trials). For the cases with the sparse and dense representations, we conducted 100 simulations for each condition, and plotted the mean±SD TD-RPE (of all the simulations, or of those corresponding to each condition (reward size) for Figure 5) at each time-step/state.

We further simulated a reward navigation task in a linear track examined in a recent study [7]. In fact, since we did not model sensory- and motor-related processes, our simulation of reward navigation task was just as same as our simulation of cue-reward association task described above, with the number of states (n) set to 5 and the reward size (R_5) set to 1 or 0.5. Since the experimental study [7] reported DA response between the "start" and "goal", which appeared to be slightly separated from the ends of the linear track (inferred from the animal's speed in Fig. 1c of [7] and also from the schematic of the linear track in Fig. 1b of [7]), we plotted TD-RPE between S_2 (named the "post-start" state) and S_4 (named the "pre-goal" state). We assumed that initially each state was represented in a punctate manner but then SR was gradually shaped through TD learning of SR features. Specifically, each state S_i (i = 1, ..., 5) was represented by a 5-dimensional feature vector:

$\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,5})$,

which was initialized as follows:

$x_{i,j} = 1$ (if $j = i$) or 0 (otherwise).

At each time-step/state i, TD error-vector of features:

$\boldsymbol{\delta}_{\mathrm{SR}i} = \boldsymbol{I}_i + \gamma\boldsymbol{x}_{i+1} - \boldsymbol{x}_i$

(if $i = 5$, the $\gamma\boldsymbol{x}_{i+1}$ term was dropped; since $\boldsymbol{x}_5$ was initialized to $\boldsymbol{I}_5$, $\boldsymbol{\delta}_{\mathrm{SR}5}$ was always $\mathbf{0}$)

was calculated, where $\boldsymbol{I}_i$ was a 5-dimensional vector whose i-th was 1 and the other elements were 0 and γ was set to 1. The feature vector $\boldsymbol{x}_i$ was updated as:

$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + a_{\mathrm{SR}}\boldsymbol{\delta}_{\mathrm{SR}i}$,

where a_{SR} was the learning rate for feature update and was set to 0.001. With these gradually updated state representations, calculation of state values and update of value-weight vector $\boldsymbol{w}$, including its initialization to $\boldsymbol{w_0} = (0.5, ..., 0.5)$ and decay, were conducted in the same manner as in the case with (constant) SR described above. a_+ and a_- were set to 0.15 and 0.075, respectively, and κ was set to 0.01. The number of trials was set to 360, which was assumed to divided into 9 sessions with each session consisting of 40 trials.

Hierarchical RL model

We considered a hierarchical RL agent having coupled two learning systems, named the circuit M and circuit L, which modeled the CDh-rmSNc and CDt-clSNc circuits, respectively (Fig. 2A). We returned to the separate punctate representation of states. Each of these two circuits had the predicted value of each state, $V_M(S_i)$ and $V_L(S_i)$ (i = 1, ..., n(= 5)) initialized to all 0. At each time-step/state, TD-RPE in each system was calculated as follows:

$\delta_{Mi} = R_i + \gamma V_M(S_{i+1}) - V_M(S_i)$ (if $i = n$, the $\gamma V_M(S_{i+1})$ term was dropped), and

$\delta_{Li} = \gamma(0.5V_L(S_{i+1}) + 0.5V_M(S_{i+1})) - V_L(S_i)$ (if $i = n$, the $\gamma(0.5V_L(S_{i+1}) + 0.5V_M(S_{i+1}))$ term was dropped)

Notably, TD-RPE in the circuit L did not contain the reward term (R_i) and instead contained the term for the predicted value of the upcoming state in the circuit M, $V_M(S_{i+1})$, which represented the conditioned reinforcement. Using these TD-RPEs, the predicted values in both circuits were updated as

$V_M(S_i) \leftarrow V_M(S_i) + a_M\delta_{Mi}$, and

$V_L(S_i) \leftarrow V_L(S_i) + a_L\delta_{Li}$,

where a_M and a_L were the learning rates for the two circuits and were set to 0.25 and 0.025, respectively. At each time-step/state during the task trials, the predicted values of all the states in the circuit M decayed as

$V_M(S_i) \leftarrow (1 - \kappa_M)V_M(S_i)$ (i = 1, ..., n),

where κ_M was the decay rate and was set to 0.01, while the predicted values in the circuit L did not decay. We also considered agent in which the $- V_L(S_i)$ term in the calculation of δ_{Li} was replaced with $- 0.5V_L(S_i)$ (Fig. 2C).

We simulated the behavior of this hierarchical RL agent in the cue-reward association learning task (200 trials, or 400 trials only for Figure 2E), and then continuously in the subsequent rest period and the extinction task. The rest period was assumed to have the same length as 200 trials, and the predicted values of all the states in the circuit M decayed 5 (number of time-steps/states) × 200 (number of trials) times according to the abovementioned equation, while the predicted values in the circuit L did not change. The extinction task (200 trials) was simulated in the same manner as the learning task except that the reward term (R_i) was set to all 0.

Distributional RL algorithms

We considered two distributional RL algorithms named the integRPE and segreRPE algorithms, which were extended from previously proposed models [27,29]. In both algorithms, there were two sets of predicted values of the states, $V_1(S_i)$ and $V_2(S_i)$ (i = 1, ..., n), initialized to all 0. $V_1(S_i)$ and $V_2(S_i)$ were presumed to be represented by D1- and D2-receptor expressing direct and indirect-pathway striatal projection neurons (SPNs), respectively, and referred to as D1 and D2 values in the Results. In reality, D2 SPNs supposedly represent values in the sign-reversed manner (i.e., higher activity for lower predicted reward value), as suggested in experimental results [29]. Also, in order to allow changes of the values to both positive and negative directions under the constraint that the weights (strengths) of corticostriatal synapses should be non-negative,

certain offsets of synaptic strengths / neural activity need to be presumed, in a similar manner to what we considered in our model for feature-based state representation in the above. However, here we did not model these biological details.

In the integRPE algorithm, at each time-step/state, a unified TD-RPE:

$\delta_i = R_i + \gamma(V_1(S_{i+1}) + V_2(S_{i+1})) - (V_1(S_i) + V_2(S_i))$ (if $i = n$, the middle term was dropped)

was calculated, where γ was set to 1. $V_1(S_i)$ and $V_2(S_i)$ were then updated as

$V_j(S_i) \leftarrow V_j(S_i) + a_{j+}\delta_i$ if $\delta_i \geq 0$ and

$V_j(S_i) \leftarrow V_j(S_i) + a_{j-}\delta_i$ if $\delta_i < 0$ ($j = 1,2$),

where the learning rates (a_{1+}, a_{1-}, a_{2+}, a_{2-}) were set to (0.15 0.05 0.05 0.15). At each time-step/state during the task trials, the two sets of predicted values of all the states decayed as

$V_j(S_i) \leftarrow (1 - \kappa)V_j(S_i)$ ($j = 1,2$; $i = 1, ..., n$),

where κ was the decay rate and was set to 0 or 0.01. This algorithm is a TD version of the AU model [27].

In the segreRPE algorithm, at each time-step/state, segregated TD-RPEs:

$\delta_{ji} = R_i + \gamma V_j(S_{i+1}) - V_j(S_i)$ ($j = 1,2$; if $i = n$, the middle term was dropped)

were calculated, where γ was set to 1. $V_1(S_i)$ and $V_2(S_i)$ were then updated as

$V_j(S_i) \leftarrow V_j(S_i) + a_{j+}\delta_{ji}$ if $\delta_{ji} \geq 0$ and

$V_j(S_i) \leftarrow V_j(S_i) + a_{j-}\delta_{ji}$ if $\delta_{ji} < 0$ ($j = 1,2$),

where the learning rates (a_{1+}, a_{1-}, a_{2+}, a_{2-}) were set to (0.15 0.05 0.05 0.15). Value-decay was not assumed in most simulations, except for those for Figure 3Bb-right, where value-decay (rate 0.01) was assumed in the same manner as in the integRPE algorithm. This segreRPE algorithm is a simplified TD version of the REDRL model [29]. The simplification that we made was to ignore the heterogeneities within D1 and D2 SPN populations. Specifically, in the REDRL simulation [29], each of D1 and D2 populations consisted of five units with $a_+/(a_+ + a_-) = 0.95, 0.85, ..., 0.55$ and $0.45, 0.35, ..., 0.05$, respectively, whereas in our segreRPE algorithm, single D1 and D2 populations had $a_+/(a_+ + a_-) = 0.75$ and 0.25, respectively. Also notably, the REDRL model first considered both asymmetric D1 and D2 plasticity for positive vs negative change in DA neuronal firing and asymmetric input-firing slope of individual DA neurons for positive vs negative RPE discovered experimentally [30] and then combined these two asymmetries into integrated parameters (corresponding to a_+ and a_- above). Therefore, in order to interpret our segreRPE algorithm as an simplification of the REDRL model, a_+ and a_- should be similarly interpreted as integrated parameters, although in the Results we only mentioned the D1 vs D2 distinction in the introduction of this algorithm.

We simulated the behavior of these algorithms in probabilistic or deterministic cue-reward association learning tasks (reward size: 1 or 3 with equal probabilities; 0 or 4 with equal probabilities; or 2 (100%)). Number of trials was set to 200 or 2000, and 100 simulations were conducted for each condition.

We quantified the strength of reward distribution- and mean-coding of the two algorithms in reference to the recent work [29], through conducting a separate set of simulations assuming some noise. Specifically, for each of the integRPE algorithm with value-decay (decay rate 0.01) and the segreRPE algorithm without value-decay, we conducted 1000 simulations for each of three conditions (reward size: 0 (100%); 1 or 3 with equal probabilities; or 2 (100%)). We considered that the results of i-th simulations (i =

1, ..., 1000) of the three conditions virtually correspond to the results of i-th pairwise recording from a D1 SPN and a D2 SPN in the three conditions (although in fact i-th simulations of the three conditions were just conducted independently and there was nothing in common). We repeated these entire simulations twice with assuming different levels of noise: pseudo-Gaussian random numbers of SD = 0.05 or 0.07 were added to D1 and D2 values at each time-step after these values for all time-steps were generated. As a measure of the strength of reward distribution-coding, we calculated an average cosine similarity of the difference in [D1 value, D2 value] vectors, at each time-step, between the condition where reward was always 2 ("fixed") and the condition where reward was 1 or 3 with equal probabilities ("variable"). We calculated 1000 difference vectors using single simulations of the two conditions (i.e., i-th simulation of reward 2 − i-th simulation of reward 1 or 3 for i = 1, ..., 1000) for each time-step. We then calculated the mean±SD of cosine similarity of each possible pair (in total 1000×999/2 = 499500 pairs) of these 1000 vectors (for each time-step). As a measure of the strength of reward mean-coding, we calculated R^2 of correlation between the reward mean and normalized noise-added D1 value and normalized noise-added D2 value (normalized, separately for D1 value and D2 value, across the three conditions for each of 1000 simulations).

Software

Simulations were conducted by using MATLAB, and pseudo-random numbers were implemented by using the functions rand, randn, and randperm.

Code Accessibility

The codes used in the present work will be deposited in GitHub upon acceptance and publication of this manuscript.

References

1. Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci *16*, 1936-1947.
2. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593-1599.
3. Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature *500*, 575-579. 10.1038/nature12475.
4. 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. 10.1038/nn.4173.
5. 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.e1625. 10.1016/j.cell.2020.11.013.
6. de Jong, J.W., Liang, Y., Verharen, J.P.H., Fraser, K.M., and Lammel, S. (2024). State and rate-of-change encoding in parallel mesoaccumbal dopamine pathways. Nat Neurosci *27*, 309-318. 10.1038/s41593-023-01547-6.
7. Guru, A., Seo, C., Post, R.J., Kullakanda, D.S., Schaffer, J.A., and Warden, M.R. (2020). Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. bioRxiv *https://doi.org/10.1101/2020.05.21.108886*.
8. Gershman, S.J. (2014). Dopamine ramps are a consequence of reward prediction errors. Neural Comput *26*, 467-471. 10.1162/NECO_a_00559.
9. Morita, K., and Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits *8*, 36. 10.3389/fncir.2014.00036.
10. Lloyd, K., and Dayan, P. (2015). Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Phasic Dopamine Signals in the Accumbens. PLoS Comput Biol *11*, e1004622. 10.1371/journal.pcbi.1004622.
11. Kato, A., and Morita, K. (2016). Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. PLoS Comput Biol *12*, e1005145. 10.1371/journal.pcbi.1005145.
12. Mikhael, J.G., Kim, H.R., Uchida, N., and Gershman, S.J. (2022). The role of state uncertainty in the dynamics of dopamine. Curr Biol *32*, 1077-1087.e1079. 10.1016/j.cub.2022.01.025.
13. Reynolds, J.N., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67-70. 35092560 [pii] 10.1038/35092560.
14. Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science *310*, 1337-1340. 310/5752/1337 [pii] 10.1126/science.1115270.
15. Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science *345*, 1616-1620. 10.1126/science.1255514.
16. 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. 10.1016/j.celrep.2022.111470.
17. Richards, B.A., and Frankland, P.W. (2017). The Persistence and Transience of Memory. Neuron *94*, 1071-1084. 10.1016/j.neuron.2017.04.037.
18. Davis, R.L., and Zhong, Y. (2017). The Biology of Forgetting-A Perspective. Neuron *95*, 490-503.

10.1016/j.neuron.2017.05.039.
19. Li, H.L., and van Rossum, M.C. (2020). Energy efficient synaptic plasticity. Elife *9*, 50804. 10.7554/eLife.50804.
20. Ryan, T.J., and Frankland, P.W. (2022). Forgetting as a form of adaptive engram cell plasticity. Nat Rev Neurosci *23*, 173-186. 10.1038/s41583-021-00548-3.
21. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (Second Edition) (MIT Press).
22. Brea, J., Urbanczik, R., and Senn, W. (2014). A normative theory of forgetting: lessons from the fruit fly. PLoS Comput Biol *10*, e1003640. 10.1371/journal.pcbi.1003640.
23. Moens, V., and Zénon, A. (2019). Learning and forgetting using reinforced Bayesian change detection. PLoS Comput Biol *15*, e1006713. 10.1371/journal.pcbi.1006713.
24. Jiang, J., Foyard, E., and van Rossum, M.C.W. (2024). Reinforcement learning when your life depends on it: A neuro-economic theory of learning. PLoS Comput Biol *20*, e1012554. 10.1371/journal.pcbi.1012554.
25. Ito, M., and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci *29*, 9861-9874. 10.1523/JNEUROSCI.6157-08.2009.
26. Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J Neurosci *35*, 8145-8157. 10.1523/JNEUROSCI.2978-14.2015.
27. Mikhael, J.G., and Bogacz, R. (2016). Learning Reward Uncertainty in the Basal Ganglia. PLoS Comput Biol *12*, e1005062. 10.1371/journal.pcbi.1005062.
28. Pinto, S.R., and Uchida, N. (2023). Tonic dopamine and biases in value learning linked through a 1 biologically inspired reinforcement learning model. bioRxiv *https://doi.org/10.1101/2023.11.10.566580*. https://doi.org/10.1101/2023.11.10.566580.
29. Lowet, A.S., Zheng, Q., Meng, M., Matias, S., Drugowitsch, J., and Uchida, N. (2024). An opponent striatal circuit for distributional reinforcement learning. bioRxiv *https://doi.org/10.1101/2024.01.02.573966*. https://doi.org/10.1101/2024.01.02.573966.
30. 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. 10.1038/s41586-019-1924-6.
31. Lowet, A.S., Zheng, Q., Matias, S., Drugowitsch, J., and Uchida, N. (2020). Distributional Reinforcement Learning in the Brain. Trends Neurosci *43*, 980-997. 10.1016/j.tins.2020.09.004.
32. Bellemare, M.G., Dabney, W., and Rowland, M. (2023). Distributional Reinforcement Learning (MIT Press).
33. Muller, T.H., Butler, J.L., Veselic, S., Miranda, B., Wallis, J.D., Dayan, P., Behrens, T.E.J., Kurth-Nelson, Z., and Kennerley, S.W. (2024). Distributional reinforcement learning in prefrontal cortex. Nat Neurosci *27*, 403-408. 10.1038/s41593-023-01535-w.
34. Morita, K., and Kato, A. (2022). Dopamine ramps for accurate value learning under uncertainty. Trends Neurosci *45*, 254-256. 10.1016/j.tins.2022.01.008.
35. Shuai, Y., Lu, B., Hu, Y., Wang, L., Sun, K., and Zhong, Y. (2010). Forgetting is regulated through Rac activity in Drosophila. Cell *140*, 579-589. 10.1016/j.cell.2009.12.044.
36. Berry, J.A., Cervantes-Sandoval, I., Nicholas, E.P., and Davis, R.L. (2012). Dopamine is required for learning and forgetting in Drosophila. Neuron *74*, 530-542. 10.1016/j.neuron.2012.04.007.
37. Hayashi-Takagi, A., Yagishita, S., Nakamura, M., Shirai, F., Wu, Y.I., Loshbaugh, A.L., Kuhlman, B., Hahn, K.M., and Kasai, H. (2015). Labelling and optical erasure of synaptic memory traces in the

motor cortex. Nature *525*, 333-338. 10.1038/nature15257.
38. Berry, J.A., Guhle, D.C., and Davis, R.L. (2024). Active forgetting and neuropsychiatric diseases. Mol Psychiatry *29*, 2810-2820. 10.1038/s41380-024-02521-9.
39. Cervantes-Sandoval, I., Chakraborty, M., MacMullen, C., and Davis, R.L. (2016). Scribble Scaffolds a Signalosome for Active Forgetting. Neuron *90*, 1230-1242. 10.1016/j.neuron.2016.05.010.
40. Castillo Díaz, F., Hernandez, M.A., Capellá, T., and Medina, J.H. (2019). Dopamine Neurotransmission in the Ventral Tegmental Area Promotes Active Forgetting of Cocaine-Associated Memory. Mol Neurobiol *56*, 6206-6217. 10.1007/s12035-019-1516-3.
41. Gallo, F.T., Zanoni Saad, M.B., Silva, A., Morici, J.F., Miranda, M., Anderson, M.C., Weisstaub, N.V., and Bekinschtein, P. (2022). Dopamine Modulates Adaptive Forgetting in Medial Prefrontal Cortex. J Neurosci *42*, 6620-6636. 10.1523/JNEUROSCI.0740-21.2022.
42. Tu, G., Ying, L., Ye, L., Zhao, J., Liu, N., Li, J., Liu, Y., Zhu, M., Wu, Y., Xiao, B., et al. (2019). Dopamine D1 and D2 receptors differentially regulate Rac1 and Cdc42 signaling in the nucleus accumbens to modulate behavioral and structural plasticity after repeated methamphetamine treatment. Biol Psychiatry *86*, 820-835. 10.1016/j.biopsych.2019.03.966.
43. Kim, H.F., and Hikosaka, O. (2013). Distinct basal ganglia circuits controlling behaviors guided by flexible and stable values. Neuron *79*, 1001-1010. 10.1016/j.neuron.2013.06.044.
44. Kim, H.F., Ghazizadeh, A., and Hikosaka, O. (2015). Dopamine Neurons Encoding Long-Term Memory of Object Value for Habitual Behavior. Cell *163*, 1165-1175. 10.1016/j.cell.2015.10.063.
45. Sharpe, M.J., Chang, C.Y., Liu, M.A., Batchelor, H.M., Mueller, L.E., Jones, J.L., Niv, Y., and Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nat Neurosci *20*, 735-742. 10.1038/nn.4538.
46. Langdon, A.J., Sharpe, M.J., Schoenbaum, G., and Niv, Y. (2018). Model-based predictions for dopamine. Curr Opin Neurobiol *49*, 1-7. 10.1016/j.conb.2017.10.006.
47. Krausz, T.A., Comrie, A.E., Kahn, A.E., Frank, L.M., Daw, N.D., and Berke, J.D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. Neuron *111*, 3465-3478.e3467. 10.1016/j.neuron.2023.07.017.
48. Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation *5*, 613-624.
49. Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J., and Daw, N.D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol *13*, e1005768. 10.1371/journal.pcbi.1005768.
50. Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., and Gershman, S.J. (2017). The successor representation in human reinforcement learning. Nat Hum Behav *1*, 680-692. 10.1038/s41562-017-0180-8.
51. Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat Neurosci *20*, 1643-1653. 10.1038/nn.4650.
52. Haruno, M., and Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. Neural Netw *19*, 1242-1254. 10.1016/j.neunet.2006.06.007.
53. Ito, M., and Doya, K. (2010). Hierarchical information coding in the striatum during decision making tasks. Neuroscience Research *68S*, e187.
54. Botvinick, M.M. (2012). Hierarchical reinforcement learning and decision making. Curr Opin Neurobiol *22*, 956-962. 10.1016/j.conb.2012.05.008.
55. Frank, M.J., and Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in

corticostriatal circuits 1: computational analysis. Cereb Cortex *22*, 509-526. 10.1093/cercor/bhr114.
56. Keramati, M., and Gutkin, B. (2013). Imbalanced decision hierarchy in addicts emerging from drug-hijacked dopamine spiraling circuit. PLoS One *8*, e61489. 10.1371/journal.pone.0061489.
57. Balleine, B.W., Dezfouli, A., Ito, M., and Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. Current Opinion in Behavioral Sciences *5*, 1-7.
58. Collins, A.G., and Frank, M.J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev *121*, 337-366. 10.1037/a0037015.
59. Lee, S.J., Lodder, B., Chen, Y., Patriarchi, T., Tian, L., and Sabatini, B.L. (2021). Cell-type-specific asynchronous modulation of PKA by dopamine in learning. Nature *590*, 451-456. 10.1038/s41586-020-03050-5.
60. Barth, A.L., and Poulet, J.F. (2012). Experimental evidence for sparse firing in the neocortex. Trends Neurosci *35*, 345-355. 10.1016/j.tins.2012.03.008.
61. Garvert, M.M., Dolan, R.J., and Behrens, T.E. (2017). A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife *6*, e17086. 10.7554/eLife.17086.
62. Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J., and Daw, N.D. (2021). Neural evidence for the successor representation in choice evaluation. bioRxiv *https://doi.org/10.1101/2021.08.29.458114*. https://doi.org/10.1101/2021.08.29.458114.
63. Bailey, D., and Mattar, M.G. (2022). Predecessor Features. arXiv *https://doi.org/10.48550/arXiv.2206.00303*. https://doi.org/10.48550/arXiv.2206.00303.
64. Haber, S.N., Fudge, J.L., and McFarland, N.R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J Neurosci *20*, 2369-2382.
65. Yin, H.H., Ostlund, S.B., and Balleine, B.W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. Eur J Neurosci *28*, 1437-1448. 10.1111/j.1460-9568.2008.06422.x.
66. Iino, Y., Sawada, T., Yamaguchi, K., Tajiri, M., Ishii, S., Kasai, H., and Yagishita, S. (2020). Dopamine D2 receptors in discrimination learning and spine enlargement. Nature *579*, 555-560. 10.1038/s41586-020-2115-1.
67. Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M.G., and Dabney, W. (2019). Statistics and Samples in Distributional Reinforcement Learning. Proceedings of the 36th International Conference on Machine Learning (eds. Chaudhuri, K. & Salakhutdinov, R.) (PMLR, 09–15 Jun 2019) *97*, 5528–5536.
68. Hamid, A.A., Frank, M.J., and Moore, C.I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. Cell *184*, 2733-2749.e2716. 10.1016/j.cell.2021.03.046.
69. Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw *15*, 535-547. S0893-6080(02)00047-3 [pii].
70. O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science *304*, 452-454. 304/5669/452 [pii] 10.1126/science.1094285.
71. Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front Neurosci *2*, 86-99. 10.3389/neuro.01.014.2008.
72. Fraser, K.M., Pribut, H.J., Janak, P.H., and Keiflin, R. (2023). From Prediction to Action: Dissociable Roles of Ventral Tegmental Area and Substantia Nigra Dopamine Neurons in Instrumental Reinforcement. J Neurosci *43*, 3895-3908. 10.1523/JNEUROSCI.0028-23.2023.

73. Masset, P., Tano, P., Kim, H.R., Malik, A.N., Pouget, A., and Uchida, N. (2023). Multi-timescale reinforcement learning in the brain. bioRxiv *https://doi.org/10.1101/2023.11.12.566754*.
74. Mohebi, A., Wei, W., Pelattini, L., Kim, K., and Berke, J.D. (2024). Dopamine transients follow a striatal gradient of reward time horizons. Nat Neurosci. 10.1038/s41593-023-01566-3.
75. Mirenowicz, J., and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature *379*, 449-451. 10.1038/379449a0.
76. Kobayashi, S., and Schultz, W. (2014). Reward contexts extend dopamine signals to unrewarded stimuli. Curr Biol *24*, 56-62. 10.1016/j.cub.2013.10.061.
77. Matsumoto, H., Tian, J., Uchida, N., and Watabe-Uchida, M. (2016). Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. Elife *5*, e17328. 10.7554/eLife.17328.
78. Lee, R.S., Sagiv, Y., Engelhard, B., Witten, I.B., and Daw, N.D. (2024). A feature-specific prediction error model explains dopaminergic heterogeneity. Nat Neurosci *27*, 1574-1586. 10.1038/s41593-024-01689-1.
79. Balleine, B.W., and O'Doherty, J.P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology *35*, 48-69. 10.1038/npp.2009.131.
80. Jiang, H., and Kim, H.F. (2018). Anatomical Inputs From the Sensory and Value Structures to the Tail of the Rat Striatum. Front Neuroanat *12*, 30. 10.3389/fnana.2018.00030.
81. Lee, K., An, S.Y., Park, J., Lee, S., and Kim, H.F. (2023). Anatomical and Functional Comparison of the Caudate Tail in Primates and the Tail of the Striatum in Rodents: Implications for Sensory Information Processing and Habitual Behavior. Mol Cells *46*, 461-469. 10.14348/molcells.2023.0051.
82. Green, I., Amo, R., and Watabe-Uchida, M. (2024). Shifting attention to orient or avoid: a unifying account of the tail of the striatum and its dopaminergic inputs. Curr Opin Behav Sci *59*. 10.1016/j.cobeha.2024.101441.
83. Menegas, W., Akiti, K., Amo, R., Uchida, N., and Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. Nat Neurosci *21*, 1421-1430. 10.1038/s41593-018-0222-1.
84. Akiti, K., Tsutsui-Kimura, I., Xie, Y., Mathis, A., Markowitz, J.E., Anyoha, R., Datta, S.R., Mathis, M.W., Uchida, N., and Watabe-Uchida, M. (2022). Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. Neuron *110*, 3789-3804.e3789. 10.1016/j.neuron.2022.08.022.
85. Tsutsui-Kimura, I., Uchida, N., and Watabe-Uchida, M. (2022). Dynamical management of potential threats regulated by dopamine and direct- and indirect-pathway neurons in the tail of the striatum. bioRxiv *https://doi.org/10.1101/2022.02.05.479267*.
86. Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci *19*, 181-189. 10.1111/j.1460-9568.2004.03095.x.
87. Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M.M., Uchida, N., and Watabe-Uchida, M. (2020). Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. Elife *9*. 10.7554/eLife.62390.
88. Matsuda, W., Furuta, T., Nakamura, K.C., Hioki, H., Fujiyama, F., Arai, R., and Kaneko, T. (2009). Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. J Neurosci *29*, 444-453. 29/2/444 [pii] 10.1523/JNEUROSCI.4029-08.2009.
89. Gangarossa, G., Espallergues, J., de Kerchove d'Exaerde, A., El Mestikawy, S., Gerfen, C.R., Hervé,

D., Girault, J.A., and Valjent, E. (2013). Distribution and compartmental organization of GABAergic medium-sized spiny neurons in the mouse nucleus accumbens. Front Neural Circuits *7*, 22. 10.3389/fncir.2013.00022.

90. Murata, K., Kanno, M., Ieki, N., Mori, K., and Yamaguchi, M. (2015). Mapping of Learned Odor-Induced Motivated Behaviors in the Mouse Olfactory Tubercle. J Neurosci *35*, 10581-10599. 10.1523/JNEUROSCI.0073-15.2015.
91. Petryszyn, S., Sánchez, M.G., Gagnon, D., Beaulieu, J.M., Parent, A., and Parent, M. (2017). A dense cluster of D_1+ cells in the mouse nucleus accumbens. Synapse *71*, 51-54. 10.1002/syn.21946.
92. Ogata, K., Kadono, F., Hirai, Y., Inoue, K.I., Takada, M., Karube, F., and Fujiyama, F. (2022). Conservation of the Direct and Indirect Pathway Dichotomy in Mouse Caudal Striatum With Uneven Distribution of Dopamine Receptor D1- and D2-Expressing Neurons. Front Neuroanat *16*, 809446. 10.3389/fnana.2022.809446.
93. Wall, N.R., De La Parra, M., Callaway, E.M., and Kreitzer, A.C. (2013). Differential innervation of direct- and indirect-pathway striatal projection neurons. Neuron *79*, 347-360. 10.1016/j.neuron.2013.05.014.
94. Lu, J., Cheng, Y., Xie, X., Woodson, K., Bonifacio, J., Disney, E., Barbee, B., Wang, X., Zaidi, M., and Wang, J. (2021). Whole-Brain Mapping of Direct Inputs to Dopamine D1 and D2 Receptor-Expressing Medium Spiny Neurons in the Posterior Dorsomedial Striatum. eNeuro *8*, ENEURO.0348-0320.2020. 10.1523/ENEURO.0348-20.2020.