Short title:

Resource allocation in dopamine neurons during reinforcement learning

Ramping and phasic dopamine activity accounts for efficient cognitive resource allocation during reinforcement learning

Minryung R. Song[1], Sang Wan Lee[1, 2, 3, 4*]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST),

[2]Program of Brain and Cognitive Engineering

[3]KAIST Institute for Health, Science, and Technology

[4]KAIST Institute for Artificial Intelligence

Daejeon, South Korea

[*]Corresponding author:

E-mail: sangwan@kaist.ac.kr (SWL)

MRS contributed to conceptualization, methodology, software, investigation, formal analysis, visualization, and writing.

SWL contributed to conceptualization, funding acquisition, project administration, resources, supervision, and writing.

Ramping and phasic dopamine activity accounts for efficient cognitive resource allocation during reinforcement learning

Abstract

Dopamine activity may transition between two patterns: phasic responses to reward-predicting cues and ramping activity arising when an agent approaches the reward. However, when and why dopamine activity transitions between these modes is not understood. We hypothesize that the transition between ramping and phasic patterns reflects resource allocation which addresses the task dimensionality problem during reinforcement learning (RL). By parsimoniously modifying a standard temporal difference (TD) learning model to accommodate a mixed presentation of both experimental and environmental stimuli, we simulated dopamine transitions and compared it with experimental data from four different studies. The results suggested that dopamine transitions from ramping to phasic patterns as the agent narrows down candidate stimuli for the task; the opposite occurs when the agent needs to re-learn candidate stimuli due to a value change. These results lend insight into how dopamine deals with the tradeoff between cognitive resource and task dimensionality during RL.

Introduction

Since cognitive and motor resources are limited, efficient resource allocation is as important as reward maximization for animals to survive. To use their resources efficiently, animals need to focus resources on the stimuli that are highly informative of reward. However, it is not easy to know which and how many stimuli in the environment are relevant to a given task. As shown in the pigeon superstition experiment by Skinner, animals can mistakenly associate reward with a stimulus that does not predict the reward [1]. Although experimentally inserted cues (e.g. tones, light) are usually more salient and informative of a reward than other stimuli in the environment (e.g. wells, floor), pseudo-conditioning has shown that animals also link the latter with the reward [2–4]. Effective task dimensions, the task dimensions to which an animal allocates resources by approaching, learning, exploiting, exploring, or paying attention to them, often differ from the essential task dimensions (Figure 1AB). Moreover, effective task dimensions may change as learning proceeds [5–7]. Efficient resource allocation would facilitate a reduction in the effective task dimensionality during RL.

Cognitive and motor resource allocation is tightly linked with an animal's approaching, avoiding, learning, exploiting, exploring, and attending; dopamine affects all of these behaviors. In the striatum, dopamine inhibition suppresses the initiation and maintenance of instrumental responses, while dopamine excitation increases the likelihood of movement initiation [8–11]. Striatal dopamine depletion reduces high-effort, high-reward choices, while increasing dopamine has the opposite effect [12–14]. Dopamine promotes avoidance responses to aversive stimuli by modulating the activity of the brainstem-projecting prefrontal neurons and nucleus accumbens neurons [15,16]. Dopamine drives both fear learning and reward learning by signaling prediction error [17–25]. Enhancing striatal dopamine increases the exploration of novel choices while decreasing the exploitation of learned options [26–28]. Prefrontal dopamine is involved in working

memory maintenance and has been suggested to reflect cognitive effort [29–31]. Dopamine neurons are necessary for the attention signal in the amygdala [32,33].

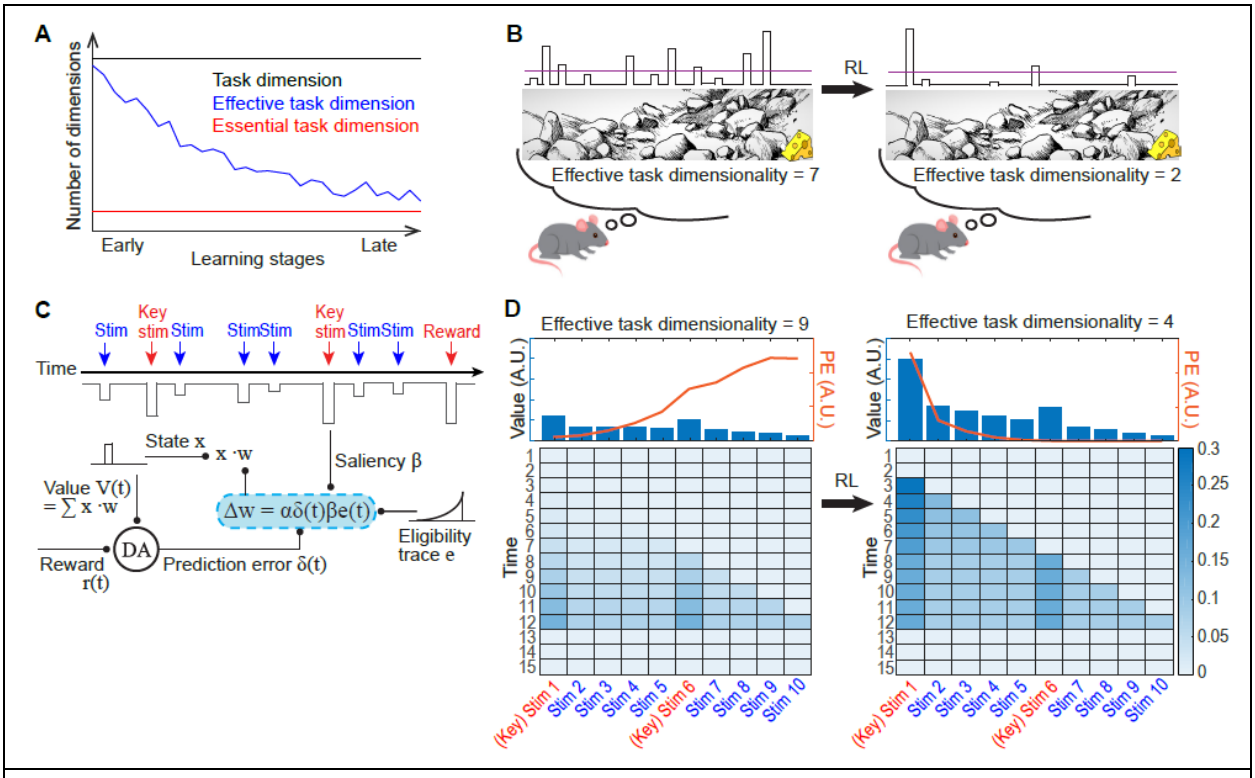

Figure 1. Resource allocating dopamine optimizes effective task dimensionality. (A) As learning proceeds, effective task dimensionality (the number of task variables affecting the behavior of the learning agent) is reduced to the minimum task dimensionality (the number of essential task dimensions). Task dimensions do not change in a fixed environment. (B) An illustration showing that effective task dimensionality decreases during RL in a stable environment. Animals do not perceive every stimulus in an environment, but they tend to recognize stimuli whose saliency exceed a threshold level (the purple horizontal line). (C) A modified TD model. Experimentally inserted cues (key stimuli) are usually more salient than environmental stimuli. (D) Value distribution and effective task dimensionality during early (left) and late (right) stages of learning. The bottom plot shows the weights of each stimulus over time. The top plot shows the prediction error and the weight of each stimulus summed across time. The number of stimuli which generate prediction error > 0.01 was considered as effective task dimensionality. Note that the main results still hold for other threshold values, such as 0.03 or 0.05.

Based on the aforementioned evidence on the involvement of dopamine in resource allocation, we assumed that dopamine activity patterns are suited to managing effective task dimensionality by efficiently allocating resources to stimuli; furthermore, we assumed that the stimuli to which resources should be assigned are primarily determined by the prediction error generated by each stimulus. The role of the dopaminergic prediction error signal in RL has been much studied both experimentally and computationally. However, investigating the role of dopamine in adjusting effective task dimensionality during RL has received scant attention, partly because most RL models have been configured to learn conditioned and unconditioned stimuli only, without accommodating the effect of other environmental stimuli on task performance. Motivated by previous findings that dopamine activity may transition between ramping to phasic [34], we hypothesized that the dopamine transition from ramping to phasic supports narrowing down candidate stimuli (reducing effective task dimensionality), whereas the opposite transition helps re-learning candidate stimuli (increasing effective task dimensionality) (Figure 1AB).

To test this hypothesis, we exploited a standard class of RL models that provide predictions of stimulus values and associated prediction errors across space and time. The prediction error signals in these models have well-simulated phasic dopamine responses to experimentally inserted cues and rewards [35–38]. RL models have also provided an explanation for the gradually increasing dopamine activity which occurs as animals approach a reward, by considering internal spatial representation, the temporal decay of dopamine-dependent synaptic potentiation, the uncertainty of action timing or discounted vigor or by assuming dopamine as a value signal [39–42]. However, no RL model has accounted for the conditions in which dopamine activity transitions between ramping and phasic. To fully accommodate a mixed presentation of both experimental and environmental stimuli, we parsimoniously modified a standard TD model and simulated dopamine activity transitions between ramping and phasic (Figure 1C). By

comparing the simulation results with experimental data from previous studies, we demonstrated that the dopamine transition supports the adjustment of effective task dimensionality during RL.

Results

The pattern of prediction error transitions from ramping to phasic as learning proceeds, reducing effective task dimensionality.

Since dopamine has a resource allocating effect in its target regions, we assumed that the stimuli that evoke sufficient dopamine excitation constitute effective task dimensions. In our model, the prediction error signal simulates dopamine activity and the number of stimuli that generate a prediction error signal larger than a threshold corresponds to effective task dimensionality. To test if effective task dimensionality decreases as learning proceeds, we considered a situation in which both environmental stimuli and experimental stimuli (key stimuli in Figure 1C) were present. In this situation, frequent exposure to stimuli would shorten the effective time window during which previously experienced stimuli affected learning. To reflect this effect, a medium λ of 0.5 was used during the simulation [43–45].

The simulation shows how the shape of the prediction error signal changes over time. During the early stages of learning, values and prediction errors were widely distributed over different stimuli (Figure 1D left). In this condition, the effective task dimensionality was high and the shape of prediction error resembled the ramping pattern of dopamine activity. As learning continues, experimental cues gained substantially higher values than other environmental stimuli (Figure 1D right). The prediction error signal gradually propagated backward and was concentrated on the initial experimental cue, exhibiting a pattern similar to the phasic dopamine activity (Figures 1D and 2A). As the shape of the prediction error transitioned from ramping to phasic, effective task dimensionality decreased (Figures 1D and 2B). The simulation results suggest that during the early stages of learning, when the learning agent usually finds it difficult to identify the key stimuli,

the agent temporarily exhibits ramping dopamine activity; this wide distribution of dopamine activity favors broad distributions of cognitive resources over multiple stimuli. As the agent gradually learns to identify reward-predicting stimuli, however, dopamine activity transitions from ramping to phasic, indicating that the agent focuses its resources on learning about those key stimuli.

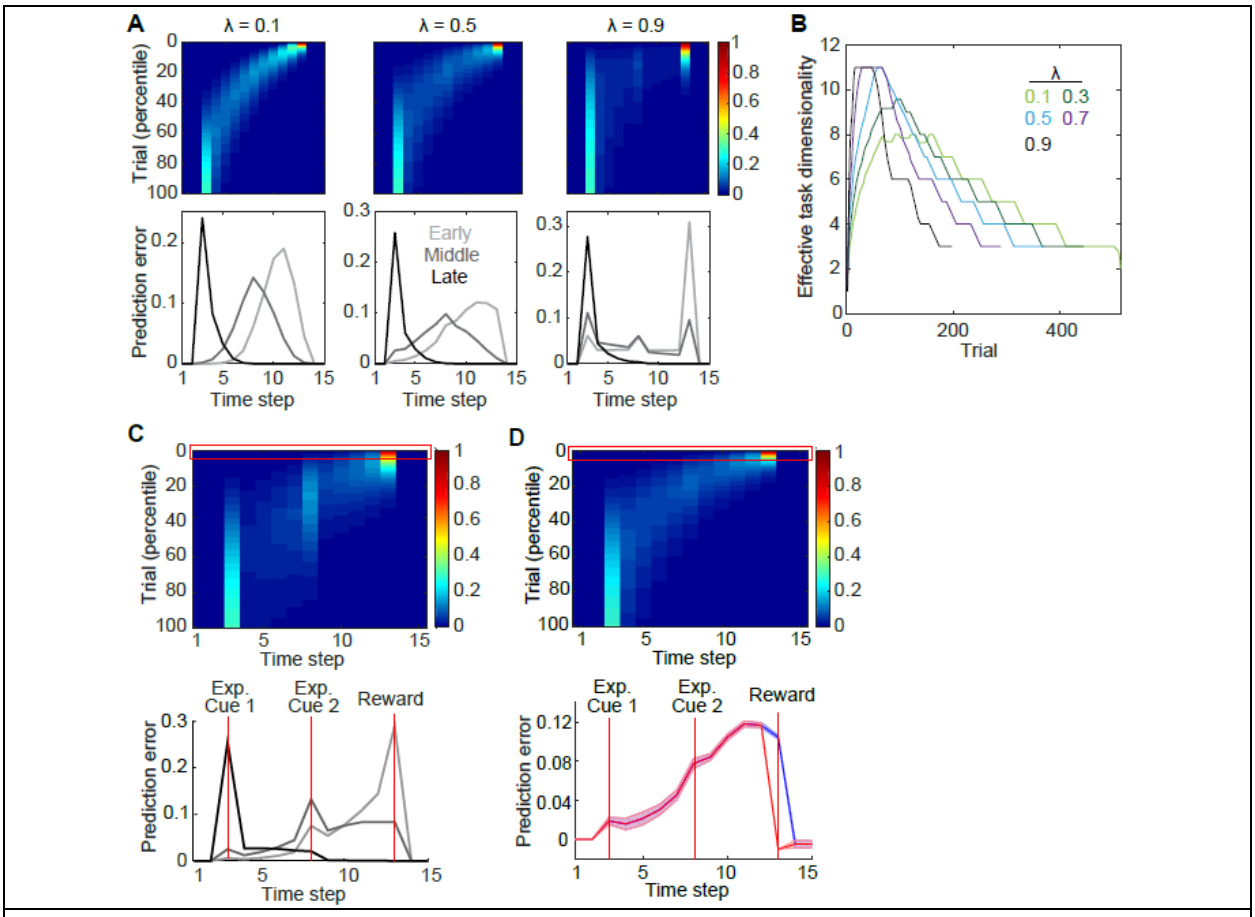

Figure 2. The ramping prediction error is caused by both the limit of cognitive resources and the environmental stimuli. (A) Prediction error signal during RL when λ is 0.1, 0.5, or 0.9. (B) Effective task dimensionality during RL. Stimuli that generated a prediction error larger than 0.01 were considered to constitute effective task dimension. Data until the agent completed learning is shown. The data were smoothed using a moving average window of 20. (C) The prediction error signal during RL when environmental stimuli were removed from the model. λ was 0.5. (D) The prediction error signal throughout

learning (top) or during the early stage of learning (bottom) when the environmental stimuli were included in the model. Dopamine concentration in the ventral striatum during rewarded (blue) or unrewarded (red) trials. We ran 100 simulations with different initial weights (random weights with a mean of 0.02). λ of 0.5 was used. Shading shows mean ± SD.

Limited resources and environmental stimuli account for the smooth, ramping shape of the prediction error.

Previous studies have shown that even in a simple RL task, limited working memory capacity influences model-free learning [43–45,47,48]. To better understand the model behavior, we ran simulations while systematically changing λ. A large λ accommodates a situation where effective task dimensionality is smaller than the working memory capacity; this occurs when the agent has a large working memory capacity or the environment is simple enough to identify task-relevant stimuli. In this scenario, the learning agent would be able to quickly pinpoint the key stimuli. Corroborating this view, for a large λ, effective task dimensionality rapidly decreases (Figure 2B); as a result, the prediction error is concentrated on the key stimuli, which produces a phasic pattern beginning in the early stages of learning (Figure 2A). This helps the agent quickly complete learning (Figure 2B).

A medium or small λ accommodates a situation in which cognitive resources can be assigned to only a portion of effective task dimensions. Reducing λ decreases the maximum size of effective task dimensionality (Figure 2B) and the width of the prediction error signal (compare the left and middle plots in Figure 2A). A wide and ramping prediction error occurs when λ has an intermediate value. The simulation results suggest that broad distributions of dopamine activity, such as ramping dopamine activity patterns, occur when effective task dimensionality is moderately smaller than the working memory capacity.

Even after the environmental stimuli are removed from the model, the prediction error shows a ramping trend if λ has an intermediate value (Figure 2C). This raises the question of whether the prediction error signal ramps up regardless of the existence of environmental cues. To answer this question, we compared our simulation results with a ramping dopamine activity reported in Howe et al. [46] (see Figure 1CE of [46]). In this study, rats were trained to travel through a large T-maze to earn a reward. The first and the second tone cues indicated the start of each trial and which arm to visit to receive the reward, respectively. A medium λ is suited to simulating this experiment because multiple environmental stimuli—such as approaching the corner of the T-maze—can provide subsidiary information to guide the animal's behavior in a large maze, increasing the effective task dimensionality. The fact that the task performance of the animals was not very high and did not reach an asymptote suggests that the data of Howe et al. [46] may have been obtained during the early stages of learning (see Figure 4E of [46]).

When λ was 0.5, both models, with and without environmental stimuli, exhibited ramping prediction error patterns during the early stage of training. While the model with environmental stimuli replicated the observed dopamine activity (Figure 2D), the model without environmental stimuli contradicted the observed dopamine activity because the prediction error inevitably peaked at the intermediate experimental cue (Figure 2C). This finding supports our hypothesis that, when the learning agent finds it difficult to identify task-relevant stimuli, dopamine ramps up, broadly distributing resources to many candidate stimuli including experimental and environmental cues.

Extended training transforms dopamine activity from ramping into phasic.

To further validate our claim, we compared the ramping-to-phasic prediction error transition from our simulation with previous experimental results. Unlike Howe et al. [46], Collins et al. [34]

trained rats for two more days after their performance reached an asymptote. In their study, the animals were trained to press two different levers to collect a reward. The authors observed that dopamine activity ramped up during early training but peaked at the first cue after extended training (Figure 3AB left). This finding is consistent with our simulation result (Figure 3AB right).

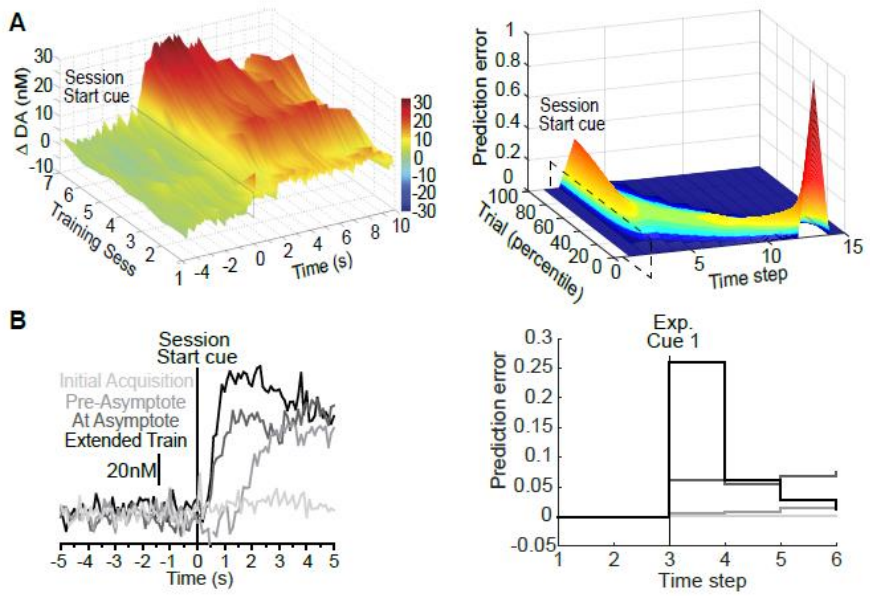

Figure 3. Dopamine transitions from ramping to phasic as learning proceeds. The left plots show the dopamine concentration in the ventral striatum observed in Collins et al. [34]. The right plots show the prediction error signal of the model. In (B), the 1st, 10th, 30th and 80th percentiles were considered as the initial acquisition, pre-asymptote, at asymptote, and extended training, respectively. The 30th percentile was chosen as the asymptote because the prediction error of the model peaks at the first cue around the 30th percentile. The parameter values were the same as in Figure 2E. Plots on the left were originally published in Collins et al. (2016) [34] under a CC-BY 4.0 license and are adapted here with permission.

Since Collins et al. [34] used a lever-press experiment while Howe et al. [46] utilized a maze paradigm, it is uncertain if extended training in a maze paradigm would transform dopamine activity from ramping into phasic. To test this possibility, we compared predictions of our model

with unpublished data collected from a maze task (see Acknowledgement). In this experiment, rats freely chose between two arms of a modified T-maze (Figure 4AB inset). The reward probability of one arm was higher than that of the other, and these probability values remained constant within a block of 17-72 trials (mean ± SD: 44.1 ± 12.0). The arm-reward relationship was reversed across four blocks without any sensory cues indicating this change. To control the inter-trial interval, a connecting bridge was lowered 2 s after the animal arrived at the entrance of the bridge (red in Figure 4B inset). Therefore, the lowering of the bridge functioned as a trial initiation cue. Note that although this study recorded electrophysiological activity of dopamine neurons while Howe et al. [46] and Collins et al. [34] used fast-scan cyclic voltammetry (FSCV) in the ventral striatum, previous studies have shown that dopamine concentration kinetics in the ventral striatum is similar to the kinetics of electrophysiological dopamine activity [10,49–51].

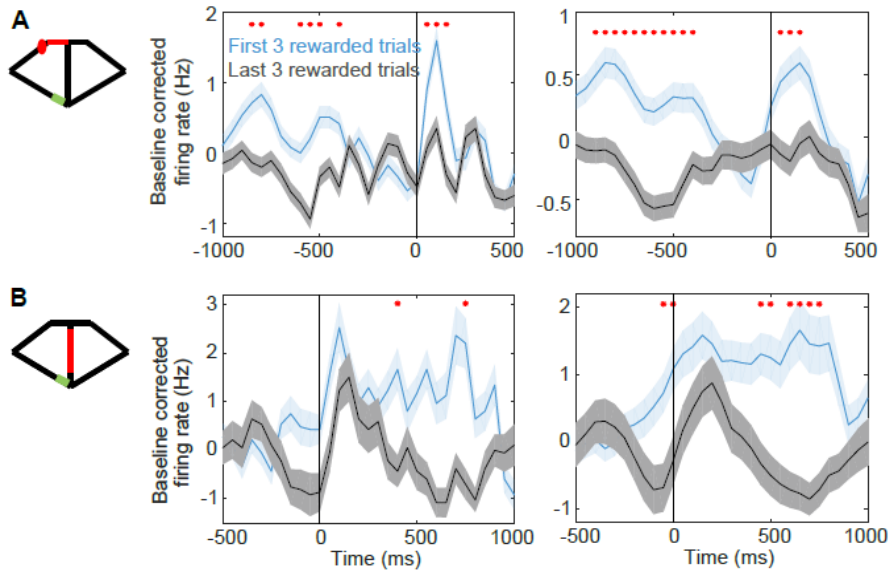

Figure 4. The dopamine response to the earlier cue becomes more phasic after extended training. (A) Dopamine response to the reward onset at the beginning and at the end of each block (4 rats, 62 dopamine cells). The inset shows the experimental apparatus. Firing rates in the red region of the inset were normalized by subtracting the mean firing rate in the green region. The reward was delivered in the region marked with a red circle in the inset. (B) Dopamine response to the bridge lowering. First 3

rewarded (light blue) and last 3 rewarded (gray) trials of the experiment are shown. The right plots show smoothed dopamine response on the left plots. A moving average window of 5 was used. Red asterisks indicate significant differences between the light blue and gray graphs (paired t-test; $p < 0.05$). Shading shows mean ± SEM.

As learning proceeded, the dopamine response to the reward onset diminished (Figure 4A). Specifically, as the dopamine activity between the bridge lowering and the reward onset decreased, the dopamine response to the bridge lowering became more phasic (Figure 4AB). These findings are consistent with the prediction of our model that the prediction error gradually backpropagates to the earlier cues and the shape of the prediction error transitions from ramping to phasic during RL. While the dopamine excitation at the reward onset decreased towards the end of each block, about four blocks were needed for the dopamine activity after the bridge lowering to significantly decrease. This is consistent with the results of Collins et al. [34], a recent experimental finding [52], and our simulation that a large phasic response to the initial cue necessitates extended training. Therefore, our model clearly replicates key features of the dopamine activity transitions from ramping to phasic.

The transition from phasic to ramping patterns upon changes in reward value.

According to our model simulation (Figure 3AB), prediction errors around the reward onset diminish as learning proceeds; this indicates a reduced sensitivity to the change in the reward value. Figure 5A shows that the prediction error at the time of reward delivery is larger when the size of the reward is doubled during early training (when the prediction error signal exhibits a ramping pattern) than late training (when the prediction error signal exhibits a phasic pattern). This simulation result indicates that extended training transforms dopamine activity from a ramping pattern to a phasic excitation to the initial cue. Further, it suggests that phasic dopamine

focuses cognitive resources on the initial cue at the expense of reduced sensitivity to environmental changes [53]. It is consistent with previous findings that extended training makes learned responses habitual and less flexible [54,55]. Recent studies have found that dopamine activity is closely linked to initiating a movement or a sequence of learned actions [8,9,19,23,24,52,56–58]. These findings, along with our simulation, suggest that during early training when an animal is uncertain about the key stimuli for the given task, many effective task dimensions—each generating a weak prediction error signal and affecting animal's behavior for a short while—guide the animal to complete the task. However, after extended training converts dopamine activity into a phasic pattern, the initial cue automatically triggers the whole behavior sequence required to complete the task [34,59].

Since phasic activity locked to a particular stimulus is not suitable for fast adaptation to environmental changes, we predicted that, when there is a change in the reward structure, the ramping pattern would reappear, facilitating re-identification of reward-predicting stimuli. Both previous experimental results and our simulation results support this prediction. In particular, our simulation (Figure 5B right) replicated the findings of Collins et al. [34] that ramping activity reappeared when the reward value was doubled after extended training (Figure 5B left; [34]). A similar effect was found in the unpublished data (Figure 5C). After the reward probabilities of the left and right arms were reversed, dopamine activity before the reward onset increased. Overall, our results show that transitions from phasic to ramping activity reflect widespread resource distribution, which allow rapid adaptation to potential changes in the stimulus value. Recent studies have found that the dopamine neurons are responsive to novel stimuli [60,61], changes to reward features [62,63], and stimuli that are potentially relevant to a given task [2–4,64–66]. These findings, together with our simulation, suggest that dopamine activity is suited to adjusting effective task dimensions in response to environmental changes.

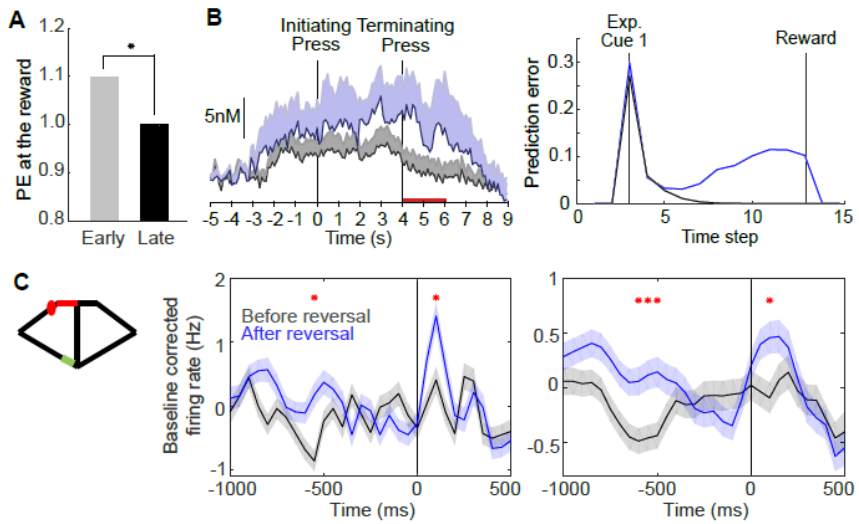

Figure 5. Dopamine transitions from phasic to ramping upon value changes. (A) Prediction error at reward delivery when the size of the reward was doubled during early or late training. We ran 100 simulations with different initial weights (random weights with a mean of 0.02). In each simulation, the first prediction error after a reward increase was calculated. The black asterisk indicates a significant difference (unpaired t-test; $p \approx 0$) (B). The experimental results of Collins et al. [34] (left) and the model behavior (right). After extended training, the ramping activity disappeared (gray). The ramping activity reappeared when the size of the reward was doubled (blue) after extended training in Collins et al. [34] (left) or during late training in the model (right). The left plot was originally published in Collins et al. (2016) [34] under a CC-BY 4.0 license. It's adapted here with permission. The right plot shows prediction error during 10th percentile of training after the reward increase. We ran 100 simulations with different initial weights. Shading shows mean ± SD. (C) Dopamine activity before and after reversal (4 rats, 62 dopamine cells). Red asterisks indicate significant differences between the blue and black graphs. (paired t-test; $p < 0.05$). Shading shows mean ± SEM.

Discussion

In the present study, we tested the theoretical possibility that dopamine transitions between ramping and phasic patterns during RL reflect efficient resource allocation. Both the simulation

and experimental results support the view that dopamine activity transitions from ramping to phasic as the RL agent narrows down the candidate reward-predicting stimuli to decrease the effective task dimensionality. The opposite occurred when the agent had to re-identify task-relevant stimuli by increasing the effective task dimensionality. This affords insight into a more fundamental question: how do animals resolve the tradeoff between prediction performance and efficient resource management?

Our model provides a testable prediction that the magnitude of the prediction error protrusion at the intermediate cue depends on the saliency contrast between experimental and non-experimental cues (Figure 6A). Our model also predicts that, the more salient a task-irrelevant stimulus, the more likely a learning agent would mistakenly associate the stimulus with a reward. Previous studies have suggested that the saliency of a stimulus is influenced by multiple factors including the stimulus' value, a sudden change in the value, uncertainty, or sensory intensity [5,33,67–71]. Since we successfully accommodated the effect of saliency by making parsimonious modifications to a standard TD model, our theory warrants the replication of a wide range of experimental data.

Moreover, our model can also be applied to a simple classical conditioning experiment, such as Pan et al. [36]. In that experiment, rats were placed in a small, simple chamber and two consecutive tone cues deterministically predicted a liquid reward. Animals only had to lick a spout to obtain the reward. This situation corresponds to the case with a large λ because effective task dimensionality is likely to be smaller than the working memory capacity (refer to Section *Limited resources and environmental stimuli…*). During early training, dopamine neurons showed strong phasic excitation to the reward, whereas the strong dopamine excitation was transferred to the initial cue during late training. Regardless of the learning stage, the omission of the second experimental cue resulted in a larger phasic response to the reward (see Figure 6B of [36]). All the dopamine activity patterns were successfully simulated by our model (Figure 6BC).

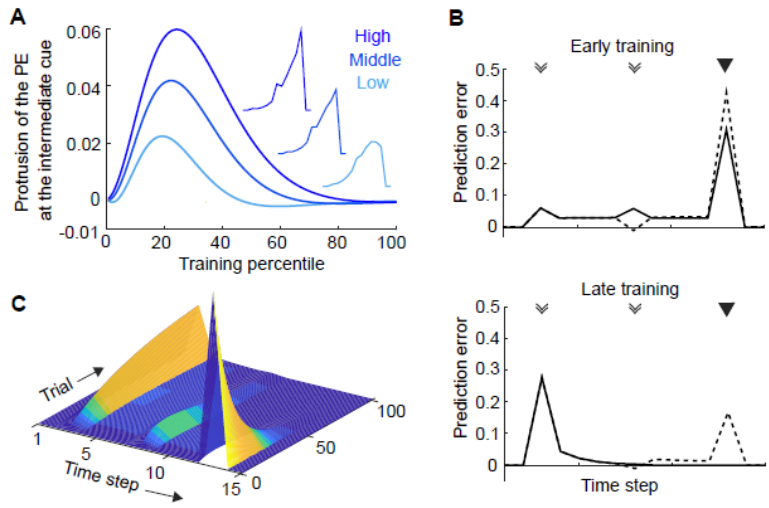

Figure 6. Model prediction and application to simple conditioning experiments. (A) The model predicted that as the saliency contrast between experimental and non-experimental cues increase, the prediction error at the intermediate experimental cue would protrude accordingly. The saliency of the experimental cues was 2, 4, and 8 for high, middle, and low contrast, respectively, while the saliency of the non-experimental (environmental) cues was 1. The level of the protrusion of the prediction error at the intermediate experimental cue was measured as the prediction error at the intermediate experimental cue minus the average of the prediction error immediately before and after the experimental cue. The inset shows the prediction error trajectory during early training. λ of 0.5 was used. (B) The simulation results of the model during early and late training. (C) The prediction error signal throughout a simple conditioning experiment with two experimental cues. λ was 0.9 in (B) and (C).

We chose the value of λ based on the presumed ratio between effective task dimensionality and the working memory capacity. According to this logic, λ should increase as learning proceeds because effective task dimensionality decreases during RL. However, we did not implement this effect and fixed the value of λ throughout each simulation for model simplicity. Adding this effect would have accelerated learning but would not have changed the major findings of the present study.

Recent empirical studies have found an increasingly diverse repertoire of dopamine activity that cannot be interpreted as a prediction error signal [17,58–63,67–70]. Our hypothesis examining the role of dopamine in resource allocation provides a framework to understand the diverse dopamine activities; prediction error is a major (but not the only) factor that determines resource allocation.

Overall, the present study provides a potential explanation for how resource allocating dopamine deals with the task dimensionality problem that RL has been known to be notoriously inefficient at tackling. Further works should investigate more fundamental problems, including the role of dopamine in resolving the tradeoff between reward maximization and resource consumption minimization.

Materials and Methods

To investigate if dopamine adjusts effective task dimensionality during RL, we incorporated environmental stimuli into a standard TD model with an eligibility trace (Figure 1C). The TD model with an eligibility trace has been shown to well account for dopamine activity during RL [36,37,52]. The goal of a TD learning agent is to maximize the expected amount of future reward. Stimulus value estimation is performed by minimizing the prediction error δ:

$$\delta(t) = r(t) + \gamma\hat{V}(t) - \hat{V}(t-1), \qquad (1)$$

where r(t) is the reward delivered at time *t* and γ is a discounting factor (0 < γ < 1) that decreases the value of delayed rewards. Here, γ was 0.9 in all simulations, and negative prediction errors were bounded by -0.01 [36]. V(t) is a value function that represents the expected value of the temporally discounted sum of all future rewards:

$$\mathrm{V(t)} = \mathrm{E}[\gamma^0 r(t) \ + \ \gamma^1 r(t+1) \ + \ \gamma^2 r(t+2) \ + \cdots]. \qquad (2)$$

Each stimulus k contributes to the estimate of *V(t)*, and the future reward estimated from each stimulus k is an inner product of the respective state vector $x_k(t)$ and the weight vector $w_k(t)$:

$$\hat{V}(\mathrm{t}) = \sum_k x_k(t) \cdot w_k(t). \quad (3)$$

If stimulus k occurs at time n, (n + m)-th element of $x_k(n + m)$ is 1 (where *m* = 0, 1, 2, 3…) and all other elements of x_k are 0. The state vector x_k enables the stimulus *k* that occurs at time n to influence the estimate of $V(n + m)$. The weights are updated as follows:

$$\Delta w_k(\mathrm{t}) = \alpha\delta(\mathrm{t})e_k(t), \quad (4)$$

where e_k is the eligibility trace for stimulus k [36] and α is the learning rate (0 < α ≤ 1). In all simulations, the learning rate α was fixed at 0.005. The eligibility trace is associated with working memory capacity [43–45]. It quantifies the degree of influence of a prediction error at time t on the value updates in previous time steps and is defined as follows:

$$e_k(t+1) = \lambda e_k(t) + x_k(t), \quad (5)$$

where λ is the eligibility trace parameter (0 ≤ λ ≤ 1). λ = 1 indicates an infinite working memory capacity; the current prediction error updates the values of all previous stimuli. A small λ indicates a small working memory capacity. Thus, using a small λ accommodates a situation in which effective task dimensionality is relatively larger than the agent's memory capacity (i.e. when many stimuli are being considered for the task). To prevent the prediction errors from carrying over to subsequent trials, the eligibility trace was reset for each trial.

Although experimentally inserted cues (e.g. tones, light) are usually more salient than others (e.g. wells, floor), pseudo-conditioning or generalization indicates that animals also associate the latter with the reward [2–4]. Previous studies have suggested that the more salient a stimulus, the more readily it should be learned [5,69,75]. In the Pearce-Hall model, the constant representing the intrinsic saliency (e.g. intensity) of the cue controls the learning rate [5,68]. To accommodate this in the current study, a saliency signal was incorporated into Equation (4) as follows:

$$\Delta w_k(t) = \alpha\delta(t)e_k(t)\beta_k, \quad (6)$$

where β_k denotes the saliency of the cues. In all simulations, β of experimental and non-experimental stimuli were 2 and 1, respectively. The agent was assumed to complete training when the value of the first experimental cue converged. To clearly demonstrate how the shape of the prediction error signal changed as learning proceeded, the 20th percentile of training—when the prediction error peaked at the second cue—was defined as the middle stage and the last 20th percentile was used to simulate the late stage of learning. Accordingly, the halfth of the middle stage, the first 10th percentile of training, was used to simulate the early stage of learning.

Acknowledgment

We thank Dr. Min Whan Jung and Dr. Sue-Hee Huh for their generosity in allowing us to use their unpublished data for this paper. We also thank Jung Hwan Shin for pre-processing the unpublished data.

Competing interests

The authors declared that no competing interests exist.

References

1. Skinner B. Superstition in the pigeon. J Exp Psychol. 1948;38: 168–172. doi:10.1037/h0055873
2. Sheafor PJ. “Pseudoconditioned” jaw movements of the rabbit reflect associations conditioned to contextual background cues. J Exp Psychol Anim Behav Process. 1975;1: 245–60. Available: http://www.ncbi.nlm.nih.gov/pubmed/1185111
3. Sheafor PJ, Gormezano I. Conditioning the rabbit’s (Oryctolagus cuniculus) jaw-movement response: US magnitude effects on URs, CRs, and pseudo-CRs. J Comp Physiol Psychol. 1972;81: 449–56. Available: http://www.ncbi.nlm.nih.gov/pubmed/4649185
4. Schultz W. Dopamine signals for reward value and risk: Basic and recent data. Behav Brain Funct. 2010;6: 1–9. doi:10.1186/1744-9081-6-24
5. Nasser HM, Calu DJ, Schoenbaum G, Sharpe MJ. The dopamine prediction error: Contributions to associative models of reward learning. Front Psychol. 2017;8: 1–17. doi:10.3389/fpsyg.2017.00244

6. Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron. Elsevier Inc.; 2017;93: 451–463. doi:10.1016/j.neuron.2016.12.040

7. Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, et al. Reinforcement learning in multidimensional environments relies on attention mechanisms. J Neurosci. Society for Neuroscience; 2015;35: 8145–57. doi:10.1523/JNEUROSCI.2978-14.2015

8. Da Silva JA, Tecuapetla F, Paixão V, Costa RM. Dopamine neuron activity before action initiation gates and invigorates future movements. Nature. Nature Publishing Group; 2018;554: 244–248. doi:10.1038/nature25457

9. Howe MW, Dombeck DA. Rapid signalling in distinct dopaminergic axons during locomotion and reward. Nature. Nature Publishing Group; 2016;535: 505–510. doi:10.1038/nature18942

10. Howard CD, Li H, Geddes CE, Jin X. Dynamic Nigrostriatal Dopamine Biases Action Selection. Neuron. Elsevier Inc.; 2017;93: 1436–1450.e8. doi:10.1016/j.neuron.2017.02.029

11. du Hoffmann J, Nicola SM. Activation of Dopamine Receptors in the Nucleus Accumbens Promotes Sucrose-Reinforced Cued Approach Behavior. Front Behav Neurosci. 2016;10: 1–19. doi:10.3389/fnbeh.2016.00144

12. Cousins MS, Atherton A, Turner L, Salamone JD. Nucleus accumbens dopamine depletions alter relative response allocation in a T- maze cost/ benefit task. Behav Brain Res. 1996;74: 189–197. doi:10.1016/0166-4328(95)00151-4

13. Berke JD. What does dopamine mean? Nat Neurosci. Springer US; 2018;21: 787–793. doi:10.1038/s41593-018-0152-y

14. Morita K, Kato A. Cognition and Behavior A Neural Circuit Mechanism for the Involvements of Dopamine in Effort-Related Choices: Decay of Learned Values, Secondary Effects of Depletion, and Calculation of Temporal Difference Error. eNeuro. 2018;5: 21–18. doi:10.1523/ENEURO.0021-18.2018

15. Vander Weele CM, Siciliano CA, Matthews GA, Namburi P, Izadmehr EM, Espinel IC, et al. Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli. Nature. Springer US; 2018;563: 397–401. doi:10.1038/s41586-018-0682-1

16. Wenzel JM, Oleson EB, Gove WN, Cole AB, Gyawali U, Dantrassy HM, et al. Phasic Dopamine Signals in the Nucleus Accumbens that Cause Active Avoidance Require Endocannabinoid Mobilization in the Midbrain. Curr Biol. Elsevier Ltd.; 2018;28: 1392–1404.e5. doi:10.1016/j.cub.2018.03.037

17. Jo YS, Heymann G, Zweifel LS. Dopamine Neurons Reflect the Uncertainty in Fear Generalization. Neuron. Elsevier Inc.; 2018;100: 916–925.e3. doi:10.1016/j.neuron.2018.09.028

18. Salinas-Hernández XI, Vogel P, Betz S, Kalisch R, Sigurdsson T, Duvarci S. Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. Elife. 2018;7: 1–25. doi:10.7554/eLife.38818

19. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci. Nature Publishing Group; 2013;16: 966–73. doi:10.1038/nn.3413

20. Chang CY, Esber GR, Marrero-Garcia Y, Yau H-J, Bonci A, Schoenbaum G. Brief

optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci. 2016;19: 111–6. doi:10.1038/nn.4191

21. Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL, et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nat Neurosci. 2017;20: 735–742. doi:10.1038/nn.4538

22. Hart AS, Rutledge RB, Glimcher PW, Phillips PEM. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci. 2014;34: 698–704. doi:10.1523/JNEUROSCI.2489-13.2014

23. Tian J, Huang R, Cohen JY, Osakada F, Kobak D, Machens CK, et al. Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons. Neuron. Elsevier Inc.; 2016;91: 1374–1389. doi:10.1016/j.neuron.2016.08.018

24. Eshel N, Tian J, Bukwich M, Uchida N. Dopamine neurons share common response function for reward prediction error. Nat Neurosci. 2016;19: 479–486. doi:10.1038/nn.4239

25. Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 2015;525: 243–246. doi:10.1038/nature14855

26. Kayser AS, Mitchell JM, Weinstein D, Frank MJ. Dopamine, locus of control, and the exploration-exploitation tradeoff. Neuropsychopharmacology. Nature Publishing Group; 2015;40: 454–462. doi:10.1038/npp.2014.193

27. Costa VD, Tran VL, Turchi J, Averbeck BB. Dopamine modulates novelty seeking behavior during decision making. Behav Neurosci. 2014;128: 556–566. doi:10.1037/a0037128

28. Beeler J a J, Daw N, Frazier CRM, Zhuang X. Tonic dopamine modulates exploitation of reward learning. Front Behav Neurosci. 2010;4: 1–14. doi:10.3389/fnbeh.2010.00170

29. Westbrook A, Braver TS. Dopamine Does Double Duty in Motivating Cognitive Effort. Neuron. Elsevier; 2016;89: 695–710. doi:10.1016/j.neuron.2016.07.020

30. Jacob SN, Ott T, Nieder A. Dopamine Regulates Two Classes of Primate Prefrontal Neurons That Represent Sensory Signals. J Neurosci. 2013;33: 13724–13734. doi:10.1523/JNEUROSCI.0210-13.2013

31. Durstewitz D, Seamans JK. The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. Biol Psychiatry. 2008;64: 739–49. doi:10.1016/j.biopsych.2008.05.015

32. Roesch MR, Esber GR, Li J, Daw ND, Schoenbaum G. Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. Eur J Neurosci. 2012;35: 1190–1200. doi:10.1111/j.1460-9568.2011.07986.x

33. Esber GR, Roesch MR, Bali S, Trageser J, Bissonette GB, Puche AC, et al. Attention-related pearce-kaye-hall signals in basolateral amygdala require the midbrain dopaminergic system. Biol Psychiatry. Elsevier Inc.; 2012;72: 1012–1019. doi:10.1016/j.biopsych.2012.05.023

34. Collins AL, Greenfield VY, Bye JK, Linker KE, Wang AS, Wassum KM. Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. Sci Rep. Nature Publishing Group; 2016;6: 1–15. doi:10.1038/srep20231

35. Schultz W, Dayan P, Montague R. A Neural Substrate of Prediction and Reward. Science. 1997;275: 1593–1599. doi:10.1126/science.275.5306.1593

36. Pan W-X, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci. 2005;25: 6235–42. doi:10.1523/JNEUROSCI.1478-05.2005

37. Pan W-X, Schmidt R, Wickens JR, Hyland BI. Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. J Neurosci. 2008;28: 9619–31. doi:10.1523/JNEUROSCI.0255-08.2008

38. Starkweather CK, Babayan BM, Uchida N, Gershman SJ. Dopamine reward prediction errors reflect hidden-state inference across time. Nat Neurosci. 2017;20: 581–589. doi:10.1038/nn.4520

39. Gershman SJ. Dopamine ramps are a consequence of reward prediction errors. Neural Comput. 2014;26: 467–71. doi:10.1162/NECO_a_00559

40. Morita K, Kato A. Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits. 2014;8: 36. doi:10.3389/fncir.2014.00036

41. Lloyd K, Dayan P. Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Phasic Dopamine Signals in the Accumbens. PLoS Comput Biol. 2015;11: 1–34. doi:10.1371/journal.pcbi.1004622

42. Kato A, Morita K. Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. PLoS Comput Biol. 2016;12: 1–41. doi:10.1371/journal.pcbi.1005145

43. Curtis CE, Lee D. Beyond working memory: The role of persistent activity in decision making. Trends Cogn Sci. Elsevier Ltd; 2010;14: 216–222. doi:10.1016/j.tics.2010.03.006

44. Todd MT, Niv Y, Cohen JD. Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. Advances in Neural Information Processing Systems 21 (NIPS 2008). 2008. pp. 1689--1696. Available: https://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observable-environments-through-dopaminergic-reinforcement

45. Lloyd K, Becker N, Jones MW, Bogacz R. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. Front Comput Neurosci. 2012;6: 1–10. doi:10.3389/fncom.2012.00087

46. Howe MW, Tierney PL, Sandberg SG, Phillips PEM, Graybiel AM. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature. Nature Publishing Group; 2013;500: 575–9. doi:10.1038/nature12475

47. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012;35: 1024–1035. doi:10.1111/j.1460-9568.2011.07980.x

48. Collins AGE, Ciullo B, Frank MJ, Badre D. Working Memory Load Strengthens Reward Prediction Errors. J Neurosci. 2017;37: 4332–4342. doi:10.1523/JNEUROSCI.2700-16.2017

49. Joosten R, Bonci A, de Ridder B, Stuber G, Feenstra M, Klanker M, et al. Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. Science (80-). 2008;321: 1690–2. doi:10.1126/science.1160873

50. Arbuthnott GW, Wickens J. Space, time and dopamine. Trends Neurosci. 2007;30: 62–9. doi:10.1016/j.tins.2006.12.003

51. Sugam JA, Day JJ, Wightman RM, Carelli RM. Phasic nucleus accumbens dopamine encodes risk-based decision-making behavior. Biol Psychiatry. Elsevier Inc.; 2012;71: 199–205. doi:10.1016/j.biopsych.2011.09.029

52. Coddington LT, Dudman JT. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci. Springer US; 2018;21: 1563–1573. doi:10.1038/s41593-018-0245-7

53. Desrochers TM, Amemori K, Graybiel AM. Habit Learning by Naive Macaques Is Marked by Response Sharpening of Striatal Neurons Representing the Cost and Outcome of Acquired Action Sequences. Neuron. Elsevier Inc.; 2015;87: 853–868. doi:10.1016/j.neuron.2015.07.019

54. Wickens JR, Horvitz JC, Costa RM, Killcross S. Dopaminergic mechanisms in actions and habits. J Neurosci. 2007;27: 8181–3. doi:10.1523/JNEUROSCI.1671-07.2007

55. Yin HH, Knowlton BJ. The role of the basal ganglia in habit formation. Nat Rev Neurosci. 2006;7: 464–476. doi:10.1038/nrn1919

56. Syed ECJJ, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME. Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci. 2016;19: 34–36. doi:10.1038/nn.4187

57. Jin X, Costa RM. Start/stop signals emerge in nigrostriatal circuits during sequence learning. Nature. 2010;466: 457–62. doi:10.1038/nature09263

58. Westbrook A, Frank M. Dopamine and proximity in motivation and cognitive control. Curr Opin Behav Sci. Elsevier Ltd; 2018;22: 28–34. doi:10.1016/j.cobeha.2017.12.011

59. Balleine BW, Dezfouli A, Lingawi NW. Habits as action sequences: hierarchical action control and changes in outcome value. Philos Trans R Soc B Biol Sci. 2014; doi:10.1098/rstb.2013.0482

60. Lak A, Stauffer WR, Schultz W. Dopamine neurons learn relative chosen value from probabilistic rewards. Elife. 2016;5: 1–19. doi:10.7554/eLife.18044.001

61. Menegas W, Babayan BM, Uchida N, Watabe-Uchida M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. Elife. 2017;6: 1–26. doi:10.7554/eLife.21886

62. Takahashi YK, Batchelor HM, Liu B, Khanna A, Morales M, Schoenbaum G. Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. Neuron. Elsevier Inc.; 2017;95: 1395–1405.e3. doi:10.1016/j.neuron.2017.08.025

63. Chang CY, Gardner M, Di Tillio MG, Schoenbaum G. Optogenetic Blockade of Dopamine Transients Prevents Learning Induced by Changes in Reward Features. Curr Biol. Elsevier Ltd.; 2017;27: 3480–3486.e3. doi:10.1016/j.cub.2017.09.049

64. Sadacca BF, Jones JL, Schoenbaum G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. Elife. 2016;5: 3–5. doi:10.7554/eLife.13665

65. Kim HF, Ghazizadeh A, Hikosaka O. Dopamine Neurons Encoding Long-Term Memory of Object Value for Habitual Behavior. Cell. Elsevier Inc.; 2015;163: 1165–1175. doi:10.1016/j.cell.2015.10.063

66. Sharpe MJ, Schoenbaum G. Evaluation of the hypothesis that phasic dopamine constitutes a cached-value signal. Neurobiol Learn Mem. 2017; doi:10.1016/j.nlm.2017.12.002

67. Dayan P, Kakade S, Montague PR. Learning and selective attention. Nat Neurosci. 2000;3: 1218–1223. Available: papers3://publication/uuid/19BE9471-AAC7-49D4-A3C5-5AC9A83645CF

68. Pearce JM, Hall G. A Model for Pavlovian Learning: Variations in the Effectiveness of Conditioned But Not of Unconditioned Stimuli. Psychol Rev. 1980;87: 532–552. Available: https://pdfs.semanticscholar.org/bfde/4e5ddaa6a968df9984b557baccd7cb38fb82.pdf

69. Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND. Differential roles of human striatum and amygdala in associative learning. Nat Neurosci. Nature Publishing Group; 2011;14: 1250–1252. doi:10.1038/nn.2904

70. Gluth S, Spektor MS, Rieskamp J. Value-based attentional capture affects multi-alternative decision making. Elife. 2018;7: 1–36. doi:10.7554/eLife.39659

71. Gottlieb J. Attention, Learning, and the Value of Information. Neuron. Elsevier Inc.; 2012;76: 281–295. doi:10.1016/j.neuron.2012.09.034

72. Fiorillo CD, Song MR, Yun SR. Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. J Neurosci. 2013;33: 4710–25. doi:10.1523/JNEUROSCI.3883-12.2013

73. Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, et al. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat Neurosci. 2016;19: 845–854. doi:10.1038/nn.4287

74. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Curr Biol. Elsevier Ltd.; 2017;27: 821–832. doi:10.1016/j.cub.2017.02.026

75. Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience. 2012. pp. 1040–1046. doi:10.1038/nn.3130