Neural Networks 126 (2020) 95–107

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Dynamic resource allocation during reinforcement learning accounts for ramping and phasic dopamine activity☆

Minryung R. Song [a], Sang Wan Lee [a,b,c,d,e,*]

[a] Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea
[b] Program of Brain and Cognitive Engineering, Daejeon, 34141, South Korea
[c] KAIST Institute for Health, Science, and Technology, Daejeon, 34141, South Korea
[d] KAIST Institute for Artificial Intelligence, Daejeon, 34141, South Korea
[e] KAIST Center for Neuroscience-inspired AI, Daejeon, 34141, South Korea

ARTICLE INFO

Article history:
Received 16 May 2019
Received in revised form 22 January 2020
Accepted 2 March 2020
Available online 10 March 2020

Keywords:
Prediction error
Salience
Temporal-difference learning model
Pearce-Hall model
Habit
Striatum

ABSTRACT

For an animal to learn about its environment with limited motor and cognitive resources, it should focus its resources on potentially important stimuli. However, too narrow focus is disadvantageous for adaptation to environmental changes. Midbrain dopamine neurons are excited by potentially important stimuli, such as reward-predicting or novel stimuli, and allocate resources to these stimuli by modulating how an animal approaches, exploits, explores, and attends. The current study examined the theoretical possibility that dopamine activity reflects the dynamic allocation of resources for learning. Dopamine activity may transition between two patterns: (1) phasic responses to cues and rewards, and (2) ramping activity arising as the agent approaches the reward. Phasic excitation has been explained by prediction errors generated by experimentally inserted cues. However, when and why dopamine activity transitions between the two patterns remain unknown. 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 them with experimental data from four different studies. The results suggested that dopamine transitions from ramping to phasic patterns as the agent focuses its resources on a small number of reward-predicting stimuli, thus leading to task dimensionality reduction. The opposite occurs when the agent re-distributes its resources to adapt to environmental changes, resulting in task dimensionality expansion. This research elucidates the role of dopamine in a broader context, providing a potential explanation for the diverse repertoire of dopamine activity that cannot be explained solely by prediction error.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

The environment is full of diverse stimuli. Since animals have a limited amount of motor and cognitive resources, they should focus their resources on potentially important stimuli. However, determining which environmental stimuli are relevant to a given task a priori is difficult. As shown in the pigeon superstition experiment by Skinner, animals can mistakenly associate reward with a stimulus that does not predict the reward (Skinner, 1948). Pseudo-conditioning also shows that animals may link a reward with environmental stimuli (e.g., wells, floors) that are less informative of the reward than are experimentally inserted cues (e.g., tones, lights) (Schultz, 2010; Sheafor & Gormezano, 1972; Sheafor, 1975). Effective task dimensions, the task dimensions to which an animal assigns resources by approaching, learning, or paying attention to them, often differ from essential task dimensions and change as learning proceeds (Fig. 1AB) (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Nasser, Calu, Schoenbaum, & Sharpe, 2017; Niv, et al., 2015). Thus, adjusting effective task dimensions through appropriate resource allocation would facilitate reinforcement learning (RL).

Previous studies have found that dopamine affects cognitive and motor resource allocation by modulating animals' approaching, learning, exploiting, exploring, and attending. Dopamine drives fear learning and reward learning by signaling prediction error (Chang, et al., 2016; Eshel, et al., 2015; Eshel, Tian, Bukwich, & Uchida, 2016; Hart, Rutledge, Glimcher, & Phillips, 2014; Jo, Heymann, & Zweifel, 2018; Salinas-Hernández, et al., 2018;

☆ 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.

* Corresponding author at: Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea.
E-mail address: sangwan@kaist.ac.kr (S.W. Lee).

https://doi.org/10.1016/j.neunet.2020.03.005
0893-6080/© 2020 Elsevier Ltd. All rights reserved.

Sharpe, et al., 2017; Steinberg, et al., 2013; Tian, et al., 2016). Dopamine excitation increases locomotor responses (Da Silva, Tecuapetla, Paixão, & Costa, 2018; du Hoffmann & Nicola, 2016; Howard, Li, Geddes, & Jin, 2017; Howe & Dombeck, 2016) and enhances the exploration of novel options (Beeler, Daw, Frazier, & Zhuang, 2010; Costa, Tran, Turchi, & Averbeck, 2014; Kayser, Mitchell, Weinstein, & Frank, 2015). Moreover, prefrontal dopamine is involved in working memory maintenance and cognitive effort (Durstewitz & Seamans, 2008; Jacob, Ott, & Nieder, 2013; Westbrook & Braver, 2016). Thus, compared to other stimuli in the environment, stimuli that excite dopamine neurons would be assigned more cognitive and motor resources. Because dopamine neurons are excited by potentially important stimuli, such as reward-predicting, intense, or novel stimuli, dopamine can allocate more resources to these stimuli, converting them into effective task dimensions. This raises an interesting possibility that dopamine activity during reinforcement learning reflects effective task dimensions.

We modified an RL model so that stimuli eliciting considerable dopamine activity, including those that are salient or generate prediction error, constitute effective task dimension (Fig. 1CD). Specifically, we incorporated environmental stimuli into a standard TD learning model to examine whether the prediction error signal reflects changes occurring in the effective task dimensionality during RL. The role of dopamine in adjusting effective task dimensionality through resource allocation has received scant attention partly because most RL models have been configured to learn conditioned and unconditioned stimuli (Huk & Hart, 2019) but not accommodate other environmental stimuli. With these modifications, our model is expected to account for a broad repertoire of dopamine activity, including ramping and phasic patterns.

Motivated by previous findings indicating that dopamine activity may transition between ramping and phasic patterns (Collins, et al., 2016), we hypothesized that dopamine transition from ramping to phasic reflects the narrowing down of candidate stimuli (reducing effective task dimensionality), whereas the opposite transition mirrors the re-learning of candidate stimuli (increasing effective task dimensionality). The prediction error signals in RL models have well-simulated phasic dopamine responses to experimentally inserted cues and rewards (Pan, Schmidt, Wickens, & Hyl, 2005, 2008; Schultz, Dayan, & Montague, 1997; Starkweather, Babayan, Uchida, & Gershman, 2017). Meanwhile, RL models have also explained the ramping dopamine activity by assuming dopamine as a value signal or by considering internal spatial representation, the temporal decay of dopamine-dependent synaptic potentiation, or the uncertainty of action timing or discounted vigor (Gershman, 2014; Hamid, et al., 2016; Kato & Morita, 2016; Lloyd & Dayan, 2015; Morita & Kato, 2014). However, the dopamine transitions between ramping and phasic patterns have not been demonstrated. By comparing our simulation results with ventral striatal dopamine concentration and dopamine neuronal firing data from previous studies, we demonstrated that resource allocation during RL can account for the dopamine transition between ramping and phasic patterns, thereby establishing a link between the dynamics of dopamine activity and effective task dimensionality. In addition, our results suggest that the ramping-to-phasic transition underlies habit formation, providing a computational account for the necessity of overtraining in a fixed environment for habit formation and insensitivity of habitual responses to environmental changes.

2. Materials and methods

2.1. Model structure

To investigate whether dopamine adjusts effective task dimensionality during RL, we parsimoniously modified a standard TD model with an eligibility trace in two ways (Fig. 1D). First, we incorporated environmental stimuli in addition to the experimental stimuli. Second, we inserted a saliency signal to simulate the different levels of saliency between the experimental and environmental stimuli. (The former is usually more salient than the latter in experimental settings).

The TD model with an eligibility trace has been shown to well account for dopamine activity during RL (Coddington & Dudman, 2018; Pan et al., 2005, 2008). 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:

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

Stimulus value estimation is performed by minimizing the prediction error δ: where r(t) is the reward delivered at time t and γ is a discounting factor ($0 < \gamma < 1$) that decreases the value of delayed rewards. Here, γ was 0.9 in all simulations. (See Figs. S 1–5 in supplementary material for the justification of parameter value selection). For the sake of biological plausibility (Pan et al., 2005, Fig. S2), we inserted a lower bound of −0.01 for prediction errors. We confirmed that although it slightly influences prediction error trajectory, it does not at all alter overall predictions about dopamine activity (see Fig. S1 in supplementary material). We picked −0.01 as a lower bound because the magnitude of dopamine suppression is about 4–6 times smaller than the size of dopamine excitation, and positive prediction error was usually less than 0.1 in our simulations. V(t) is a value function that represents the expected value of the temporally discounted sum of all future rewards:

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

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 $w_k(t)$:

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

If stimulus k occurs at time n, (n + m)th element of $x_k(n + m)$ is 1 (where $m = 0, 1, 2, 3\ldots$) 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:

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

where e_k is the eligibility trace for stimulus k (Pan et al., 2005) and α is the learning rate ($0 < \alpha \leq 1$). The eligibility trace is associated with working memory capacity (Curtis & Lee, 2010; Lloyd, Becker, Jones, & Bogacz, 2012; Todd, Niv, & Cohen, 2008). 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:

$$e_k(t+1) = \lambda e_k(t) + x_k(t), \tag{5}$$

where λ is the eligibility trace parameter ($0 \leq \lambda \leq 1$). $\lambda = 1$ 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), pseudoconditioning or generalization indicates that animals also

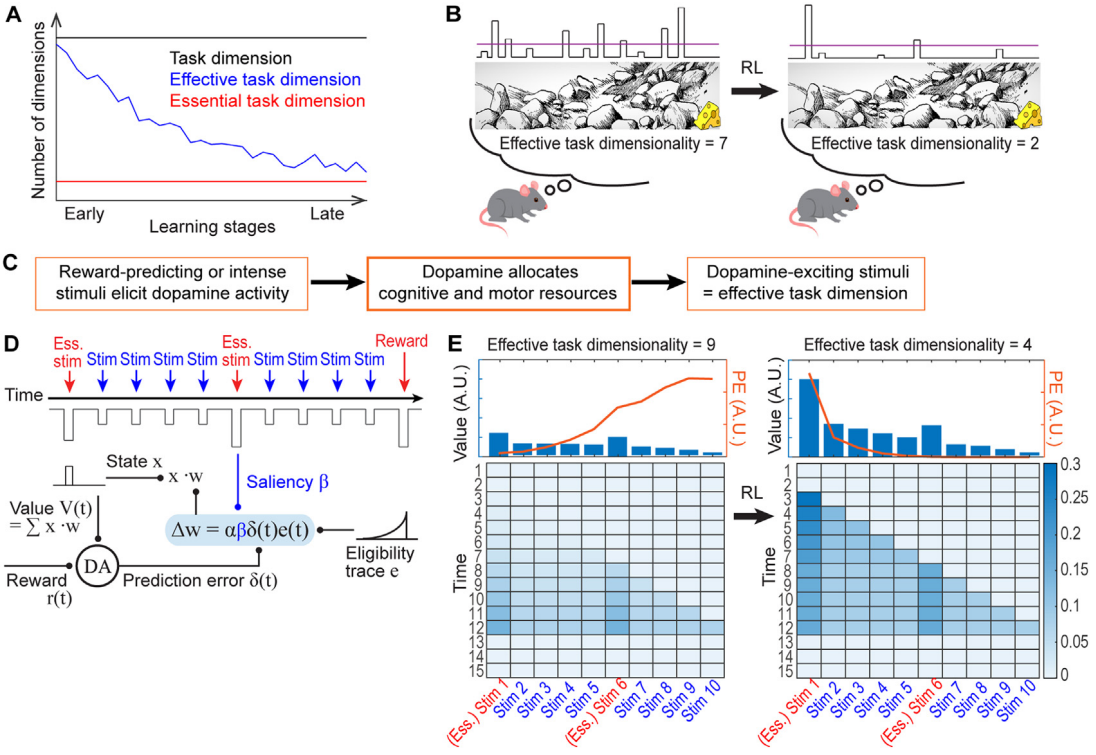

Fig. 1. By allocating resources dopamine optimizes effective task dimensionality. (A) In a stable environment, in which task dimensionality does not change, effective task dimensionality (the number of stimuli to which the learning agent assigns cognitive or motor resources to perform a task) usually decreases as learning proceeds and is lower-bounded by the number of essential task dimensions. (B) Illustration of the relationship between effective task dimensionality and resource allocation. The height of each bar represents the amount of resource allocated to each stimulus. The purple horizontal line indicates the saliency threshold for animals' recognition. (C) Because dopamine modulates resource allocation, stimuli that excite dopamine neurons – such as prediction error-generating or intense stimuli – constitute effective task dimensionality. (D) A modified TD model implementing dynamic resource allocation to environmental cues. Letters in blue show modifications to the standard TD model used. Experimentally inserted cues ("essential stimuli" in red) are usually more salient than environmental cues ("stimuli" in blue). Essential stimuli constitute an essential task dimension. Here, "time" refers to time points within trials. (E) 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 that generate a prediction error > 0.01 is considered as effective task dimensionality. Note that the main results still hold for other threshold values, such as 0.03 and 0.05.

associate the latter with the reward (Schultz, 2010; Sheafor & Gormezano, 1972; Sheafor, 1975). Previous studies have suggested that the more salient a stimulus, the more readily it should be learned (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Nassar, et al., 2012; Nasser et al., 2017). In the Pearce–Hall model, the constant representing the intrinsic saliency (e.g. intensity) of the cue controls the learning rate (Nasser et al., 2017; Pearce & Hall, 1980). To accommodate this in the current study, a saliency signal was incorporated into Eq. (4) as follows:

$$\Delta w_k(t) = \alpha \delta(t)\, e_k(t)\, \beta_k, \tag{6}$$

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 extended training. Accordingly, the halfth of the middle stage, the first 10th percentile of training, was used to simulate the early stage of learning. In all simulations, α and γ were fixed at 0.005 and 0.9, respectively. Only λ was modulated.

In order to simulate the unpublished data obtained from a two-armed bandit task with reversal, we modified the above model slightly. The animals' choice (left or right) was implemented with a softmax function as follows:

$$P_C = \frac{\exp(\tau V_C)}{\exp(\tau V_L) + \exp(\tau V_R)}, \quad where \;\; C = L \;\; or \;\; R. \tag{7}$$

Here, τ $(0 \leq \tau)$ is the temperature parameter. As τ decreases below 1, all actions become equally likely. When τ is very large, the action with a higher value is almost always selected. Previous reversal studies have shown that the absolute value of the prediction error is large at the beginning of each block, which modulates the learning rate (Esber, et al., 2012; Li et al., 2011). Based on these studies, we adjusted Eq. (4) as below:

$$\Delta w_k(t) = \alpha \rho_n \delta(t)\, e_k(t)\, \beta_k, \tag{8}$$

where ρ_n is the surprise signal for trial n, which depends on the previous trial's prediction error as follows:

$$\rho_{n+1} = \eta\, |\delta_n| + (1-\eta)\rho_n, \tag{9}$$

where η determines the effect size of the previous trial's prediction error. As η increases, the learning rate of a trial is more strongly influenced by the previous trial's prediction error. Because the positive prediction error usually ranged between 0.1 and 1 while simulating a reversal task, we adjusted the lower bound for the prediction error to -0.1. All the codes are available at https://github.com/brain-machine-intelligence/dopamine-resource-allocation/.

2.2. Experimental design and analysis of the unpublished data

Four young male Sprague Dawley rats were used. The animals were restricted to 30 min of access to water after finishing one behavioral session per day. To maintain a stable level of water deprivation, their body weights were maintained at 80%–85% of their free-feeding weights.

The experiments were conducted in the dark phase of a 12 h light/dark cycle. The experimental protocol was approved by KAIST IACUC.

The behavioral task described in Lee, Ghim, Kim, Lee, and Jung (2012) was used. The animals were trained in a dynamic two-armed bandit task in a modified T-maze (Fig. 5A). The reward probability of one arm was higher than that of the other (0.72/0.12 or 0.63/0.21), and these probability values remained constant within a block of 17–72 trials (mean ± SD: 44.1 ± 12.0). Each experiment consisted of 4 blocks (149–180 trials, mean ± SD: 175.4 ± 5.9), and each animal underwent the experiment 11–20 times (mean ± SD: 15.0 ± 3.9). The arm–reward relationship was reversed across four blocks without any sensory cue indicating this change. A connecting bridge was lowered 2 s after the animal arrived at the entrance of the bridge (purple area in Fig. 5A) to control the inter-trial interval. Therefore, the lowering of the bridge functioned as the trial initiation cue. The maze (65 × 60 cm, track width: 8 cm, wall height: 3 cm) was elevated 30 cm from the floor and contained four photobeam sensors (red lines in Fig. 5A) to monitor the animal's position. The water reward was delivered in the region marked with orange area in Fig. 5A.

Tetrodes targeting the ventral tegmental area were implanted while the animals were anesthetized with isoflurane (1.5–2.0% [vol/vol] in 100% oxygen). Putative single units were isolated using the software MClust (A.D. Redish). Only clusters with L-ratio < 0.1, isolation distance > 19, and peak-to-peak amplitude >150 μA were included in the analysis. The recorded units were classified as putative dopamine neurons, putative GABA neurons, and unidentified neurons on the basis of mean discharge rates, coefficients of variation (CV) of their inter-spike intervals, and half-valley widths of the filtered spike waveforms using a Gaussian mixture model (Shin, Kim, & Jung, 2018; Stark, Rothe, Wagner, & Scheich, 2004).

Cell firing rates in Figs. 5D–G and 6A–D were baseline corrected by subtracting the mean firing rate between −1 and 0 s before the entrance of the connecting bridge. Paired t-tests were used in Figs. 5DE and 6AB and unpaired t-tests were used in Figs. 5FG and 6CD.

3. Results

3.1. 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 (essential stimuli in Fig. 1D) were present. In this situation, frequent exposure to stimuli would shorten the effective time window during which previously experienced stimuli affect learning. To reflect this effect, a medium λ of 0.5 was used during the simulation (Curtis & Lee, 2010; Lloyd et al., 2012; Todd et al., 2008).

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 (Fig. 1E 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 (Fig. 1E 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 (Figs. 1E and 2C). As the shape of the prediction error transitioned from ramping to phasic, effective task dimensionality decreased (Figs. 1E and 2F). The simulation results suggest that during the early stages of learning, when the learning agent usually finds it difficult to identify the essential 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 essential stimuli.

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

Identifying essential task dimensions becomes increasingly difficult as there are more non-essential stimuli in the environment. Exposure to non-essential stimuli interrupts learning of the essential task dimensions, and non-essential stimuli themselves may affect dopamine activity (Schultz, 2010; Sheafor & Gormezano, 1972; Sheafor, 1975). In our model, the former is implemented by λ and the latter by prediction errors generated by non-essential stimuli. To better understand the model behavior, we first ran simulations while systematically changing λ.

A large λ accommodates a situation where effective task dimensionality is small compared to the learning agent's capacity, and the agent can quickly pinpoint the essential stimuli. Corroborating this view, for a large λ, effective task dimensionality rapidly decreases (Fig. 2F); as a result, the prediction error is concentrated on the essential stimuli, which produces a phasic pattern beginning in the early stages of learning (Fig. 2A). This helps the agent quickly complete learning (Fig. 2F).

The situation that a large λ accommodates is similar to simple classical conditioning experiments, such as that in Pan et al. (2005). In that experiment, rats were placed in a small (floor area 25 × 16.5 cm), simple chamber, and two consecutive tone cues deterministically predicted a liquid reward. Animals only had to lick a spout to obtain the reward. The simulation result produced using a large λ also resembled the typical phasic dopamine activity that has been observed in simple conditioning experiments (Coddington & Dudman, 2018; Pan et al., 2005). In Pan et al. (2005), dopamine neurons showed strong phasic excitation to the reward during early training, whereas the strong dopamine excitation was transferred to the initial cue after extended training (hundreds of trials; Fig. 2B left). Regardless of the learning stage, the omission of the second experimental cue resulted in a larger phasic response to the reward. When we used a large λ, our model successfully simulated all the dopamine activity patterns (Fig. 2B right). Note that although our model is designed to simulate dopamine concentration in the ventral striatum, previous studies have shown that dopamine concentration kinetics in the ventral striatum is comparable to that of electrophysiological dopamine activity (Arbuthnott & Wickens, 2007; Stuber, et al., 2008; Sugam, Day, Wightman, & Carelli, 2012). These simulation results suggests that phasic dopamine activity occurs when task dimensionality is smaller than the agent's cognitive capacity.

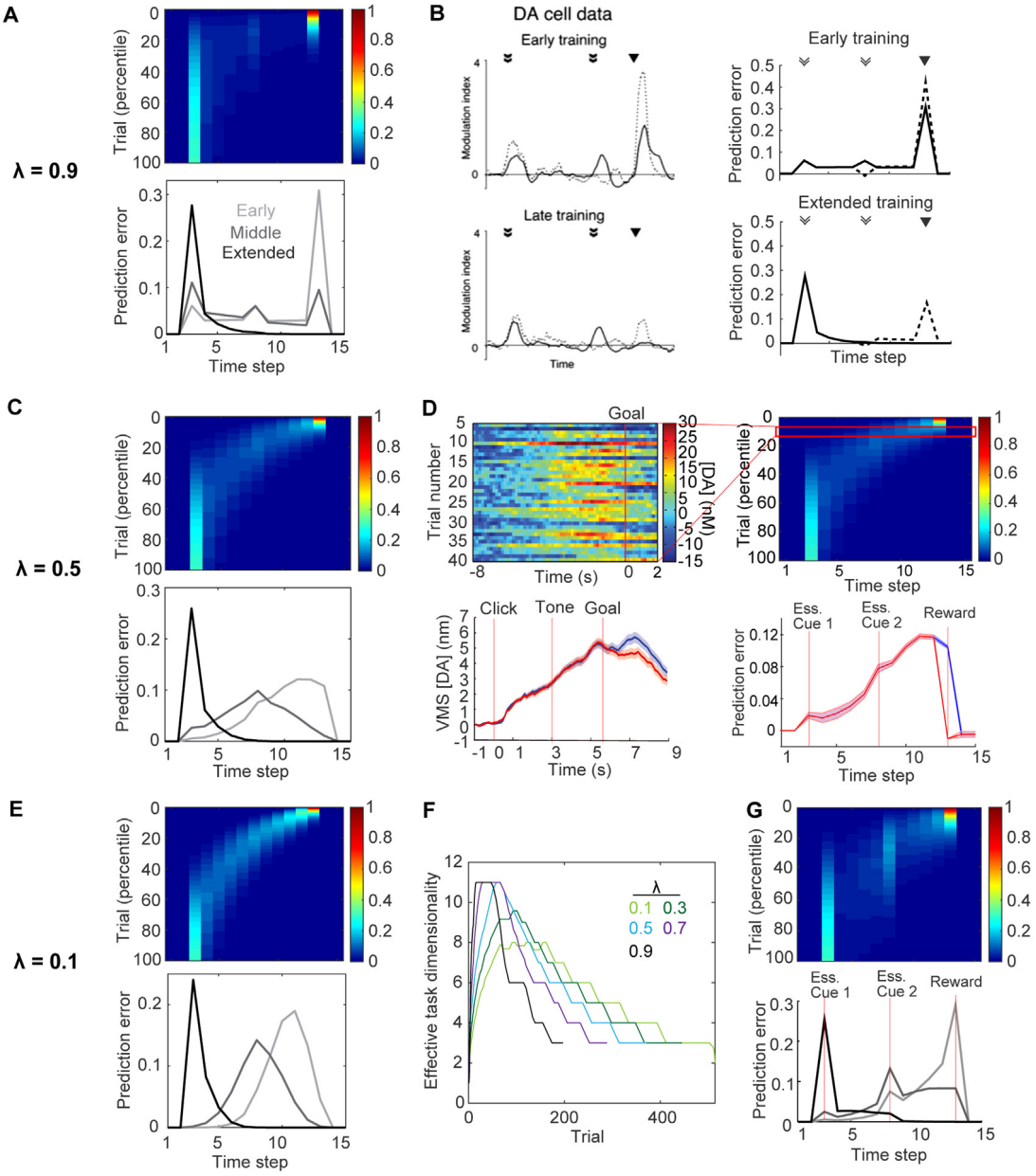

Fig. 2. The ramping prediction error is caused by both the limit of cognitive resources and the environmental stimuli. (A, C, E) Prediction error signal during RL when λ is 0.1, 0.5, or 0.9, respectively. (B) The left plots show dopamine activity during early and extended training in a simple conditioning task with two experimental cues. Two double arrows and a black triangle on top of each plot denote the timing of the two experimental cues and reward, respectively. The solid and dotted lines indicate dopamine responses when both of the cues were presented and those when the second cue was omitted, respectively. Figures were adapted from Pan et al. (2005) with permission (Copyright (2005) Society for Neuroscience). The right plots show simulation results of the model during early and extended training. λ was 0.9. (D) The left plots show dopamine concentration in the ventral striatum during rewarded (blue) or unrewarded (red) trials. The right plots show the prediction error signal throughout learning (top) or during the early stage of learning (bottom). The red box in the top plot marks early training during which the prediction error exhibits a ramping trend. λ of 0.5 was used. We ran 100 simulations with different initial weights (random weights with a mean of 0.02). Shading shows mean ± SD. The figure was adapted from Howe et al. (2013) with permission. (F) 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. (G) The prediction error signal during RL when environmental stimuli ("stim" in blue in Fig. 1D) were removed from the model. λ was 0.5.

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 (Fig. 2F) and the width of the prediction error signal (compare Fig. 2C and E). 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 task dimensionality is slightly smaller than the agent's cognitive capacity. This provides a potential explanation for why ramping dopamine activity has been rarely observed in rather simple experiments (Flagel, et al., 2011; Sugam et al., 2012) but has started being reported in recent, more sophisticated experiments (Collins, et al., 2016; Hamid, et al., 2016; Howe et al., 2013; Mohebi, et al., 2019). For example, Howe et al. (2013) (Fig. 2D left) found a ramping dopamine activity. In this study, rats were trained to travel more than a meter 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 non-essential 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. When λ was 0.5, our model replicated the observed dopamine activity (Fig. 2D right).

Next, we tested the effect of prediction errors generated by non-essential stimuli on the prediction error trajectory. Even after the environmental stimuli are removed from the model, the prediction error shows a ramping trend during early training if λ has an intermediate value (Fig. 2G). However, while the model with environmental stimuli replicated the observed dopamine activity (Fig. 2D right), the model without environmental stimuli contradicted the observed dopamine activity because the prediction error inevitably peaked at the intermediate experimental cue (Fig. 2G). This finding supports our hypothesis that, when the learning agent finds it difficult to identify essential task dimensions, dopamine ramps up, broadly distributing resources to many candidate stimuli including experimental and environmental cues.

3.3. Extended training in a fixed environment transforms dopamine activity from ramping into phasic

Our simulation result shows that, even when a medium λ is used, the prediction error trajectory transitions from ramping to phasic patterns after extended training (Fig. 2C). After several times more training than the amount the agent typically takes until the ramping pattern appears, the first experimental cue generates a large prediction error. This simulation result suggests that extended training in a fixed environment transforms dopamine activity from a ramping pattern to a phasic excitation to the initial cue.

On the other hand, extended training diminished prediction errors around the reward onset. This indicates a reduced sensitivity to the change in the reward value. Fig. 3A 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 extended training (when the prediction error signal exhibits a phasic pattern). It is consistent with previous findings that overtraining makes learned responses habitual and less sensitive to outcome (Wickens, Horvitz, Costa, & Killcross, 2007; Yin & Knowlton, 2006).

The ramping-to-phasic transition was not found in Howe et al. (2013) (Fig. 2D left). We cannot rule out the possibility that rats in Howe et al. (2013) might not have been over-trained enough to convert learned responses into automatic and almost habitual behavior. This is supported by the fact that the task performance of the animals was not very high and did not reach an asymptote (see Fig. 4E of Howe et al., 2013).

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. (2013), Collins, et al. (2016) trained rats for two more days after their performance reached an asymptote (Fig. 3B). In their study, the animals were trained to press two different levers to collect a reward, which was delivered deterministically; note that the static reward environment makes it easy for rats to be habitual. The rats were trained for two more days after the average time between the initial level press and reward collection reached an asymptote. The authors observed that dopamine activity ramped up during early training but peaked at the first cue after extended training (Fig. 3C–E top). This finding is consistent with our simulation result (Fig. 3C–E bottom). The reason why dopamine activity after extended training in Collins, et al. (2016) is not completely "phasic" is probably because the amount of training was still not sufficient. Our model simulation suggests that until dopamine activity finishes the ramping-to-phasic transition, it exhibits transitive forms, including those shown in Collins, et al. (2016) (Fig. 3C bottom). Overall, these results support our claim that training extensive enough to make learned responses automatic induces the ramping-to-phasic transition. So far, the transition from ramping to phasic activity has been rarely reported, probably because ramping activity is a relatively new discovery, and few studies have examined the effect of overtraining in a fixed environment on the ramping pattern. Our model provides a useful prediction for the further study of the ramping-to-phasic transition.

Decreasing sequence time (Fig. 3B) is consistent with our claim, too. Unlike the experimenter, animals do not know which stimuli in the environment are essential task dimensions from the start of training. During early training, they often allocate cognitive and motor resources to non-essential task dimensions (e.g., floor, walls) by sniffing, touching, or looking around in the environment. Thus, non-experimental stimuli as well as essential stimuli (here, the two levers) constitute effective task dimensions, particularly during early training. As the animal learns that non-experimental stimuli are not essential for the task, the set of its effective task dimensions would gradually reduce to the set of the essential task dimensions (Fig. 1A). Meanwhile, as the animal pays less attention to non-essential stimuli, its responses would become more efficient, thereby shortening response time, as shown in Fig. 3B.

Collins, et al. (2016) observed that although extended training developed a fast, stereotypical sequence of actions, the rats occasionally deviated from their typical performance. Nucleus accumbens dopamine concentration during those trials was more "ramping" and significantly higher than that during typical performance (Fig. 3F). This finding is consistent with our claim that ramping activity reflects resource allocation to non-essential stimuli, whereas phasic excitation to the initial cue after extended training promotes the execution of inflexible, habitual action sequences.

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

Since phasic activity locked to a particular stimulus is not suitable for fast adaptation to environmental changes (Fig. 3A), we predicted that, when there is a change in the reward structure, the ramping pattern would reappear, facilitating re-identification of reward-predicting stimuli. Indeed, dopamine activity is suited to adjusting effective task dimensions in response to environmental changes because dopamine neurons are excited by novel stimuli (Lak, Stauffer, & Schultz, 2016; Menegas, Babayan, Uchida, & Watabe-Uchida, 2017), changes to reward features (Chang, Gardner, Di Tillio, & Schoenbaum, 2017; Takahashi, et al., 2017). Both previous experimental results and our simulation results support this prediction. Collins, et al. (2016) found that ramping activity reappeared when the reward value was doubled after extended training (Fig. 4A top). Because the prediction error during the 30th percentile (Fig. 3C bottom) resembles the dopamine activity after extended training in Fig. 4A top (gray), we doubled the size of the reward at the 30th percentile of learning. After the reward increased, the prediction error displayed a phasic-to-ramping transition very similar to the transition shown in Fig. 3C bottom (Fig. 4B). Collins, et al. (2016) conducted 15 trials after the reward increase and averaged the dopamine response. The recurrent ramping activity resembled the prediction error signal during early training after the reward increase but not the signal immediately after the reward increase (Fig. 4A bottom). Overall,

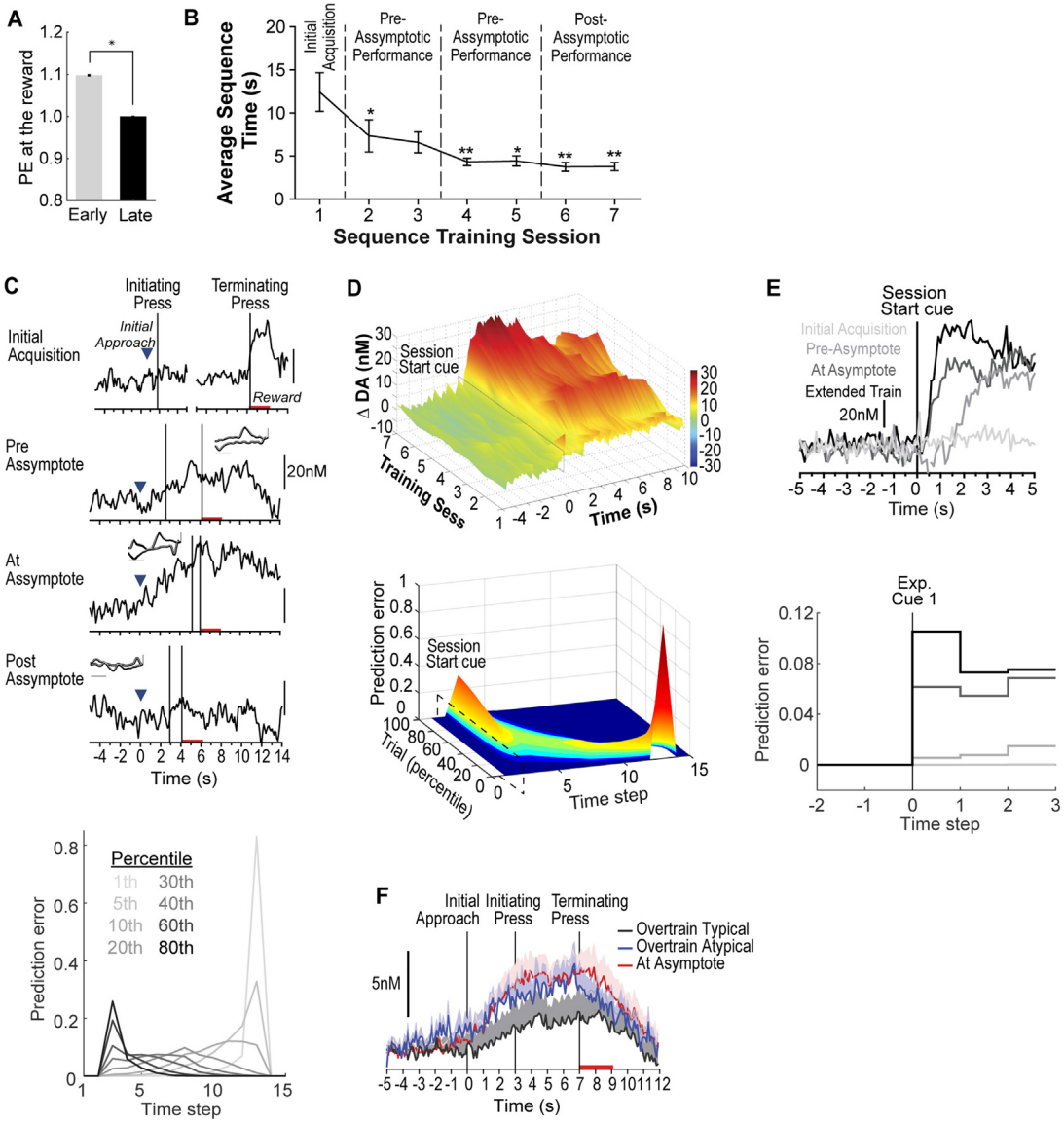

Fig. 3. Dopamine transitions from ramping to phasic as learning proceeds. (A) Prediction error at reward delivery when the size of the reward was doubled during early or late training. We ran 100 simulations with different initial weights (random weights with a mean of 0.02). The black asterisk indicates a significant difference (unpaired *t*-test; $p \approx 1.9189e{-}158$). (B) Average time between the initial lever press and reward collection across training in Collins, et al. (2016). In (C–E), the top plots show the dopamine concentration in the ventral striatum observed in Collins, et al. (2016), and the bottom plots show the prediction error signal of the model. The parameter values were the same as in Fig. 2D right. (C) A representative single trial. The bottom plot shows the shape of the prediction error signal at 1th, 5th, 10th, 20th, 30th, 40th, 60th, and 80th percentiles of training. The agent was assumed to complete training when the value of the first experimental cue converged. In (E), the 1st, 10th, 30th, and 40th percentiles were considered as the initial acquisition, pre-asymptote, at asymptote, and extended training, respectively. (F) Average dopamine concentration during asymptotic performance (red), typical performance (black), and atypical performance (blue) after extended training in Collins, et al. (2016). (B), (C–E) top plots and (D) were originally published in Collins, et al. (2016) under a CC-BY 4.0 license and adapted here with permission.

our simulation results show that, whereas extensive training in a fixed environment forms a habit and transforms dopamine activity from ramping to phasic, the ramping activity occurs in changing environments, increasing effective task dimensionality and sensitivity to environmental changes. Supporting this result, recent reversal studies (Hamid, et al., 2016; Mohebi, et al., 2019) have found ramping dopamine activity.

3.5. Dopamine transition between phasic and ramping in extended reversal training

Since Collins, et al. (2016) used a lever-press experiment while Howe et al. (2013) utilized a maze paradigm, it is uncertain if extended training in a maze paradigm would transform dopamine activity from ramping into phasic. To investigate this possibility, we tested predictions of our model using unpublished data collected from a modified T-maze task (see Acknowledgment). In this experiment, each trial begins as a rat entered the medial stem (red arrow in Fig. 5A). Two seconds after the animal entered the medial stem, a connecting bridge (purple in Fig. 5A) was lowered, allowing the animal to cross it. Therefore, the lowering of the bridge was considered as trial initiation. The animal freely chose either the left or the right arm to earn a reward. The reward probability of one arm was set to be higher than that of the other (0.72/0.12 or 0.63/0.21), and these probability values remained constant within a block. The arm–reward association was

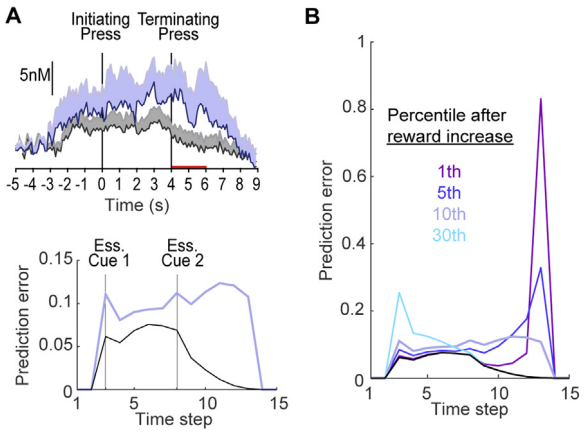

Fig. 4. Dopamine transitions from phasic to ramping upon value changes. (A) The experimental results of Collins, et al. (2016) (top) and the model behavior (bottom). The ramping activity disappeared after extended training (black) but reappeared when the size of the reward was doubled (purple). The top plot was originally published in Collins, et al. (2016) under a CC-BY 4.0 license. The parameter values were the same as in Fig. 3. Among many prediction error trajectories shown in Fig. 3C bottom, the prediction error during the 30th percentile most resembles the dopamine activity after extended training shown in Fig. 4B top (gray). The size of the reward was doubled following the 30th percentile. The prediction error during the 10th percentile after the reward increase is shown in purple. (B) The prediction error during the 1th, 5th, 10th, and 30th percentiles after the reward increase (black).

reversed across four blocks without any sensory cues indicating this change.

In order to simulate a reversal task, we modified our model so that the model chose left or right action based on a softmax function (Fig. 5B). In Fig. 5B, experimental (or essential) stimuli, non-essential stimuli, and the reward are indicated by a red-, blue-, or orange-colored number, respectively. Based on previous reversal studies showing that large prediction errors occurring at the beginning of each block enhance the learning rate (Esber, et al., 2012; Li et al., 2011), we made the prediction error generated by the previous outcome adjust the learning rate (see Section 2.2).

In this reversal experiment, the environment remains constant until the animal exits from the bridge, while the environment after the animal making its choice changes across blocks (Fig. 5A). Each animal was extensively trained (1959–3584 trials; 11–18 days) in the unpublished data. We predicted that dopamine activity during the constant part of the reversal task would gradually transition from the ramping pattern to phasic excitation to the initial cue through extended training (Fig. 3C bottom). We also predicted that dopamine activity during the changing part of the task would transition from phasic excitation by the reward to the ramping pattern, and that the size of ramping activity would gradually diminish as in Fig. 4B.

To test this hypothesis, we divided dopamine responses to the bridge lowering into those obtained during the first 9 days and the rest obtained during the last 9 days. In each set of data, dopamine activity several hundred milliseconds after the bridge lowering diminished as learning proceeded, resulting in a more "phasic" shape by the end of each block (Fig. 5DE left). When the data from the first 9 days and the last 9 days were compared (Fig. 5FG left), the transition from the ramping to phasic pattern became more pronounced. Our model replicates these features of the data (Fig. 5D–G right). In order to demonstrate how prediction error changes during reversal learning, we gathered simulation results from the first 400 rewarded trials in Fig. 5C. Fig. 5C shows that the size of the prediction error generated by the first essential cue increases as learning proceeds, whereas the prediction error between the first essential cue and the reward decreases. Taken together, the experimental data and our simulation results support our hypothesis that dopamine activity transitions from the ramping pattern to phasic excitation to the initial essential cue in a static environment.

We next examined the dopamine responses during the changing part of the task. As expected, dopamine activity before the reward onset decreased as learning proceeded (Fig. 6A–D left). Although not statistically significant, phasic excitation by the reward also diminished during learning. Our model replicated both features of the data (Fig. 6A–D right). These results are consistent with our prediction that phasic excitation to a reward that arises immediately after an environmental change gradually transitions to the ramping pattern, and that the phasic excitation and the ramping activity decrease as learning proceeds.

4. 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. These results suggest that dopamine deals with the task dimensionality problem during RL through resource allocation.

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 (Dayan, Kakade, & Montague, 2000; Esber, et al., 2012; Gluth, Spektor, & Rieskamp, 2018; Gottlieb, 2012; Li et al., 2011; Nasser et al., 2017; Pearce & Hall, 1980). Multiple neural substrates, including striatal acetylcholine, norepinephrine, the amygdala, and dopamine, have been implicated in salience signaling (Cox & Witten, 2019; Esber, et al., 2012; Likhtik & Johansen, 2019; Nasser et al., 2017). However, interaction between the neural substrates as well as how salience signaling changes during learning remain elusive. To preclude confusion arising from the as of yet poorly understood mechanism underlying this interaction, we fixed beta in all of our simulation conditions. Note that without exploring beta values for each simulation, we were able to replicate a wide range of experimental data. Moreover, our model provides a testable prediction that the prediction error at the intermediate experimental cue would protrude accordingly as the saliency contrast between experimental (essential) and non-experimental (non-essential) cues increased (Fig. 7). This prediction provides a potential explanation for why a salient but task-irrelevant stimulus, such as the lever in the Skinner's superstition experiment, can be mistakenly associated with a reward (Skinner, 1948); the more salient a stimulus is, the more strongly it excites dopamine neurons, resulting in more resources assigned to the stimulus (Steinberg, et al., 2013).

In our model, dopamine contributes to salience processing in two ways. First, the dopaminergic prediction error signal is affected by the cue saliency (β) due to its role in the value updates. Second, dopamine activity can signal salience. Although many dopamine neurons signal the reward prediction error (Coddington & Dudman, 2019; Eshel et al., 2016; Tian, et al., 2016), increasing evidence suggests that the response profiles of dopamine

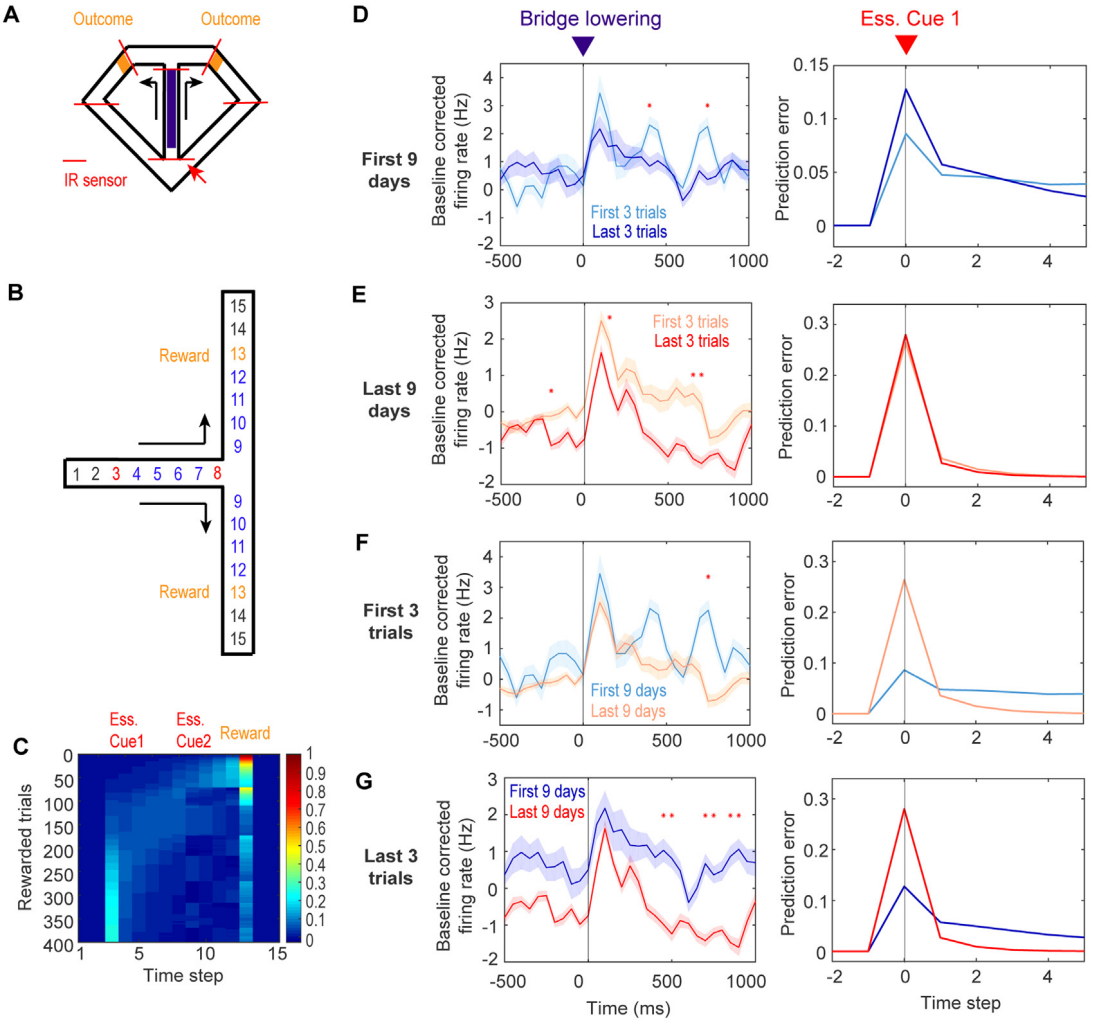

Fig. 5. Extensive reversal training turned dopamine response to the initial cue from ramping to phasic. (A) Experimental apparatus of the unpublished data. A modified T-maze was used, and infra-red sensors detected the animals' location. Red arrow marks the entrance of the medial stem. Two seconds after the animal entered the medial stem, a connecting bridge (purple) was lowered. (B) Structure of the model for the reversal task. Red, blue, and orange numbers indicate the time step when experimental (or essential) stimuli, non-essential stimuli, and the reward were presented, respectively. A left or right direction was chosen using a softmax function, intended to implement stochastic value-based decision making. The length of each block was 100 trials, and the choice for the higher reward probability was alternated across blocks. High and low reward probabilities were 100% and 0%, respectively. τ of 3 and η of 0.9 were used. The values of all other parameters were the same as in Figs. 3 and 4. (C) Prediction error signal of the model during the first 400 rewarded trials. (D–G) Dopamine response to the bridge lowering during the first 3 and last 3 rewarded trials in each block (N = 62) (left) and the model behavior (right). For (D–G) right, the first block when initial acquisition occurred was excluded. Simulation results during the first 5 and next 5 blocks after the first block were used. (D) and (E) show dopamine activity during the first 9 (N = 30) and last 9 days (N = 32), respectively. (F) and (G) were made from (D) and (E) to help compare dopamine activity during the first 9 days and last 9 days. Red asterisks indicate significant differences between the two graphs in each plot (paired t-test; $p < 0.05$). Shading shows mean ± SEM.

neurons are not uniform (Cox & Witten, 2019; Engelhard, Finkelstein, Cox, Fleming, Jang, Ornelas, et al., 2019; Lammel, et al., 2008; Lammel, Ion, Roeper, & Malenka, 2011). Previous studies have shown that some dopamine neurons respond to potentially important stimuli, such as intense, novel, or extinguished but previously reward-associated stimuli (Fiorillo, Yun, & Song, 2013; Kim, Ghazizadeh, & Hikosaka, 2015; Lak et al., 2016; Menegas et al., 2017). Both the prediction error signal modulated by salience and the sensitivity of dopamine to salient stimuli would help allocate cognitive and motor resources to potentially important stimuli. Recent empirical studies have found an increasingly diverse repertoire of dopamine activity that cannot be interpreted as a prediction error signal (Berke, 2018; Coddington & Dudman, 2019; Howe et al., 2013; Lammel et al., 2011; Lau, Monteiro, & Paton, 2017; Menegas et al., 2017; Pignatelli & Bonci, 2015; Schultz, 2016). Our hypothesis, which examines the role of dopamine in resource allocation, provides a framework for understanding the diverse activities of dopamine; prediction error is a major (but not the only) factor that determines resource allocation, and dopamine response to intense, novel, or motor response-related stimuli can be interpreted in terms of effective task dimensionality.

In our study, a moderate level of eligibility trace (medium λ) was a primary condition for the ramping pattern. The eligibility trace matters because a biological agent has limited cognitive capacity, and the environment contains many non-essential stimuli. A hypothetical agent with infinite cognitive capacity ($\lambda = 1$) would be able to quickly pinpoint the essential task dimensions by taking into account all previously encountered stimuli simultaneously regardless of the number and diversity of non-essential stimuli during learning. However, every biological agent has a limited cognitive capacity, and its learning can be retarded by non-essential stimuli consuming cognitive resources. Previous studies have shown that even in a simple RL task, working memory load influences model-free learning (Collins, Ciullo, Frank, &

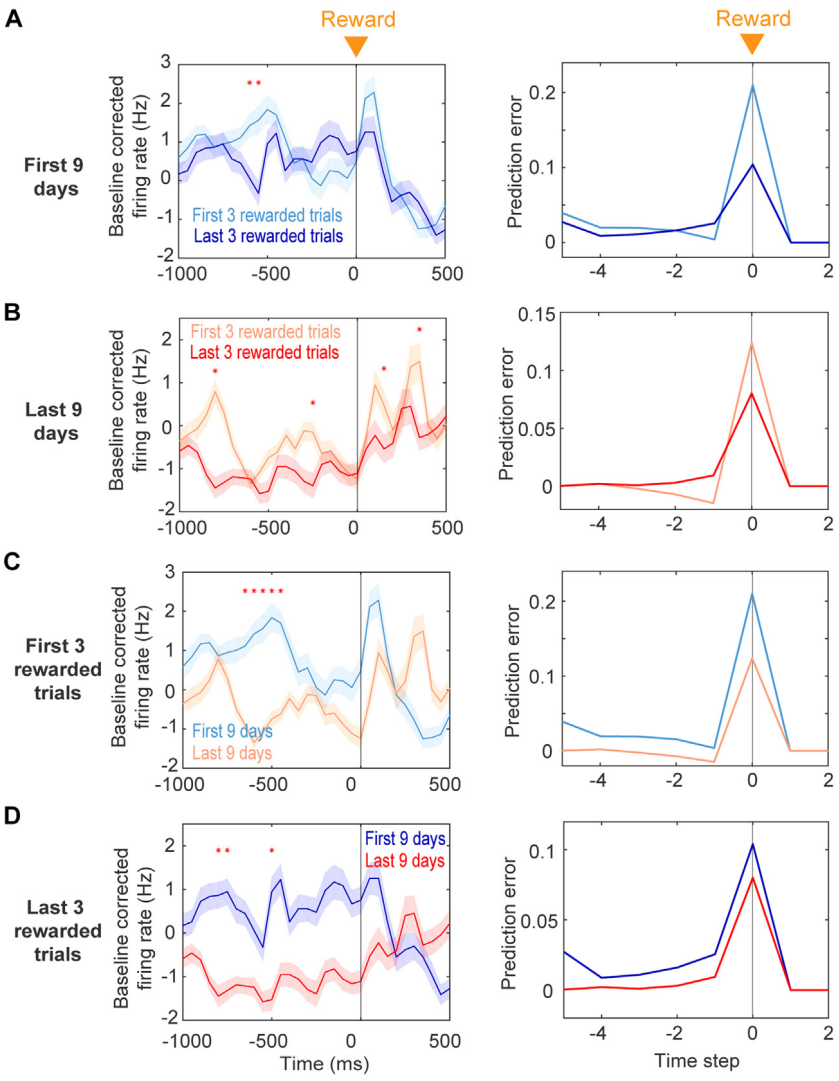

Fig. 6. Dopamine response to the reward gradually diminished during reversal training. (A–D) Dopamine response to the reward onset during the first 3 and last 3 rewarded trials in each block (N = 62) (left) and the model behavior (right). Excluding the first block when initial acquisition occurs, simulation results during the first 5 and next 5 block after the first block were used. The values of all parameters are same as Figs. 3–5. (A) and (B) show dopamine activity during the first 9 (N = 30) and last 9 days (N = 32), respectively. (C) and (D) were made from (A) and (B) to help comparing dopamine activity during the first 9 days and last 9 days. Red asterisks indicate significant differences between the two graphs in each plot (paired t-test; $p < 0.05$). Shading shows mean ± SEM.

Badre, 2017; Collins & Frank, 2012; Curtis & Lee, 2010; Lloyd et al., 2012; Todd et al., 2008). In our model with non-essential, environmental stimuli (Fig. 1D), λ allows the model to examine the effect of non-essential stimuli on RL. The influence of non-essential stimuli on learning has received scant attention in RL research. The present study proposes that non-essential stimuli are worth considering in RL models.

Non-essential stimuli in our model also helped us in simulating habit. A habit is an outcome-insensitive, chunked sequence of actions, the initial action of which empowers the execution of the full sequence (Smith & Graybiel, 2016). Converting a stimulus–response association into a habit requires much more training than just acquiring the association, and dopamine is involved in this process (Graybiel, 2008; Yin, Zhuang, & Balleine, 2006). To our knowledge, our model is the first to provide a computational account for why habit formation necessitates overtraining and why habitual responses are resistant to environmental changes. Recent studies have found that dopamine activity is closely linked to initiating a movement or a sequence of learned actions (Coddington & Dudman, 2018; Da Silva et al., 2018; Eshel et al., 2016; Howe & Dombeck, 2016; Jin & Costa, 2010; Steinberg, et al., 2013; Syed, et al., 2016; Tian, et al., 2016; Westbrook & Frank, 2018). Along with our simulation, these findings suggest that during early training, when an animal is uncertain about the essential stimuli for the given task, many effective task dimensions – each generating a weak prediction error signal and affecting the 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, which accompanies reduced sensitivity to environmental changes (Balleine, Dezfouli, & Lingawi, 2014; Collins, et al., 2016).

Extensive evidence has linked dopamine with motivation (Berke, 2018; Berridge, 2007; Bromberg-Martin, Matsumoto, & Hikosaka, 2010; Coddington & Dudman, 2019; Da Silva et al., 2018; Engelhard et al., 2019; Flagel, et al., 2011). Specifically, recent studies have suggested that the ramping pattern is related to motivation (Howe et al., 2013; Mohebi, et al., 2019). Because we defined effective task dimension as task dimensions to which the learning agent allocates cognitive and motor resources, effective task dimensionality and motivation are not mutually exclusive. The notion of effective task dimension includes motivation and

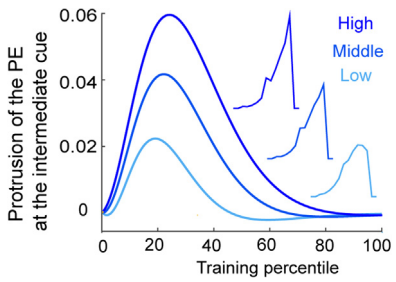

Fig. 7. Model prediction. The model predicted that the magnitude of the prediction error protrusion at the intermediate cue depends on the saliency contrast between experimental and non-experimental cues. The saliency of the experimental cues was 8, 4, and 2 for high, middle, and low contrast, respectively, while the saliency of the non-experimental (environmental) cues was 1. The values of the other parameters are the same as in Figs. 3–6. 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.

therefore has a broader application. For instance, Fig. 3F displays dopamine activity during typical and atypical performance of the animal after extended training. Dopamine concentration was higher and more "ramping" during atypical than typical action sequences. Since atypical behaviors are less efficient at earning the reward than fast, stereotypical performance, this suggests that the animals were less motivated during atypical performance. Although this finding does not fit the idea that ramping activity signals motivation to earn a reward, it is consistent with our claim that ramping activity reflects resource allocation to non-essential stimuli. Dissociating the effect of motivation on ramping activity from effective task dimensionality in computational models requires a more precise definition of task dimension and motivation, respectively. We think that this limitation affords opportunities for future work that is aimed at dissociating the effect of motivation on ramping activity from task dimensionality.

Ramping dopamine also has been proposed to signal state value (Hamid, et al., 2016). In our study, task dimensionality is claimed to be tightly linked to reward prediction error, and it is the signal that drives value updates. So it would be tricky to separate out the effective task dimensionality from value state changes. To fully address this issue, future research should consider recording dopamine activity with tasks that can separately manipulate these two variables.

Because previous studies have shown that the kinetics of dopamine concentration in the ventral striatum was comparable to that of dopamine spiking activity (Arbuthnott & Wickens, 2007; Stuber, et al., 2008; Sugam et al., 2012), we applied our model to both FSCV data obtained from the ventral striatum and electrophysiological data. However, dopamine concentration in a brain region is not determined solely by dopamine spiking activity but rather by multiple factors, such as the density of dopamine varicosities (Arbuthnott & Wickens, 2007), the mechanism of dopamine clearance (Durstewitz & Seamans, 2008), and region-specific modulation of dopamine release (Cox & Witten, 2019; Liu & Kaeser, 2019). Therefore, dopamine concentration in a brain region and dopamine cell firing might exhibit different patterns at the same time. For example, a recent study found that the dopamine concentration in the nucleus accumbens ramps while dopamine neurons fire phasically (Mohebi, et al., 2019). Our model and dopamine firing in the unpublished data showed phasic excitation to the reward immediately after environmental changes but a ramping pattern by the end of each block (Figs. 4 and 6AB). On the other hand, Collins, et al. (2016) reported a ramping dopamine concentration after the size of reward was doubled (Fig. 4A gray). This warrants further research to determine whether this discrepancy is ascribed to the effect of averaging the first 15 trials after the reward increase or to the difference between dopamine firing and its concentration. Specifically, a recent study reported dopamine neurons whose spiking activity was modulated by spatial distance from reward similar to ramping dopamine concentration in the ventral striatum (Engelhard et al., 2019). We hope future studies on dopamine activity heterogeneity and region-specific mechanisms of dopamine release would elucidate the distinction between dopamine firing and release.

5. Conclusion

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

Acknowledgments

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.

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019M3E5D2A01066267), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence) and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-06.

Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2020.03.005.

References

Arbuthnott, G. W., & Wickens, J. (2007). Space, time and dopamine. *Trends in Neurosciences*, *30*, 62–69, Available at: http://www.ncbi.nlm.nih.gov/pubmed/17173981. (Accessed 1 October 2013).

Balleine, B. W., Dezfouli, A., & Lingawi, N. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, Available at: http://dx.doi.org/10.1098/rstb.2013.0482.

Beeler, J. a. J., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience*, *4*, 1–14, Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2991243/. (Accessed 23 September 2013).

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*, 787–793, Available at: http://dx.doi.org/10.1038/s41593-018-0152-y.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl)*, *191*, 391–431, Available at: http://www.ncbi.nlm.nih.gov/pubmed/17072591. (Accessed 16 September 2013).

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, *68*, 815–834, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3032992&tool=pmcentrez&rendertype=abstract. (Accessed 17 September 2013).

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., & Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nature Neuroscience*, *19*, 111–116, Available at: http://www.ncbi.nlm.nih.gov/pubmed/26642092. (Accessed 8 December 2015).

Chang, C. Y., Gardner, M., Di Tillio, M. G., & Schoenbaum, G. (2017). Optogenetic blockade of dopamine transients prevents learning induced by changes in reward features. *Current Biology*, *27*, 3480–3486.e3. Available at: https://doi.org/10.1016/j.cub.2017.09.049.

Coddington, L. T., & Dudman, J. T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nature Neuroscience*, *21*, 1563–1573, Available at: http://dx.doi.org/10.1038/s41593-018-0245-7.

Coddington, L. T., & Dudman, J. T. (2019). Learning from action: Reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, *104*, 63–77, Available at:https://linkinghub.elsevier.com/retrieve/pii/S0896627319307421.

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*, 4332–4342, Available at: http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2700-16.2017.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024–1035.

Collins, A. L., Greenfield, V. Y., Bye, J. K., Linker, K. E., Wang, A. S., & Wassum, K. M. (2016). Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. *Scientific Reports*, *6*, 1–15, Available at: http://dx.doi.org/10.1038/srep20231.

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

Cox, J., & Witten, I. B. (2019). Striatal circuits for reward learning and decision-making. *Nature Reviews Neuroscience*, *20*, Available at:http://dx.doi.org/10.1038/s41583-019-0189-2.

Curtis, C. E., & Lee, D. (2010). Beyond working memory: The role of persistent activity in decision making. *Trends in Cognitive Sciences*, *14*, 216–222, Available at: http://dx.doi.org/10.1016/j.tics.2010.03.006.

Da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*, 244–248, Available at: http://dx.doi.org/10.1038/nature25457.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223, Available at: http://papers3://publication/uuid/19BE9471-AAC7-49D4-A3C5-5AC9A83645CF.

du Hoffmann, J., & Nicola, S. M. (2016). Activation of dopamine receptors in the nucleus accumbens promotes sucrose-reinforced cued approach behavior. *Frontiers in Behavioral Neuroscience*, *10*, 1–19, Available at: http://journal.frontiersin.org/Article/10.3389/fnbeh.2016.00144/abstract.

Durstewitz, D., & Seamans, J. K. (2008). The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. *Biological Psychiatry*, *64*, 739–749, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18620336 (Accessed 19 September 2013).

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., et al. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, *570*, 509–513, Available at: http://dx.doi.org/10.1038/s41586-019-1261-9.

Esber, G. R., Roesch, M. R., Bali, S., Trageser, J., Bissonette, G. B., Puche, A. C., et al. (2012). Attention-related pearce-kaye-hall signals in basolateral amygdala require the midbrain dopaminergic system. *Biological Psychiatry*, *72*, 1012–1019, Available at: http://dx.doi.org/10.1016/j.biopsych.2012.05.023.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., & Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, *525*, 243–246, Available at: http://www.nature.com/doifinder/10.1038/nature14855.

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience*, *19*, 479–486, Available at: http://www.nature.com/doifinder/10.1038/nn.4239.

Fiorillo, C. D., Yun, S. R., & Song, M. R. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, *33*, 4693–4709, Available at: http://www.ncbi.nlm.nih.gov/pubmed/23486943. (Accessed 1 October 2013).

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., et al. (2011). A selective role for dopamine in stimulus-reward learning. *Nature*, *469*, 53–57, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3058375&tool=pmcentrez&rendertype=abstract. (Accessed 24 January 2014).

Gershman, S. J. (2014). Dopamine ramps are a consequence of reward prediction errors. *Neural Computation*, *26*, 467–471, Available at: http://www.ncbi.nlm.nih.gov/pubmed/24320851. (Accessed 23 July 2015).

Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *Elife*, *7*, 1–36.

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron*, *76*, 281–295, Available at: http://dx.doi.org/10.1016/j.neuron.2012.09.034.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, *31*, 359–387, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18558860. (Accessed 17 September 2013).

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Weele, C. M. Vander, et al. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, *19*, 117–126, Available at: http://www.nature.com/doifinder/10.1038/nn.4173. (Accessed 23 November 2015).

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience*, *34*, 698–704, Available at: http://www.jneurosci.org/content/34/3/698.short. (Accessed 30 July 2015).

Howard, C. D., Li, H., Geddes, C. E., & Jin, X. (2017). Dynamic nigrostriatal dopamine biases action selection. *Neuron*, *93*, 1436–1450.e8. Available at: http://dx.doi.org/10.1016/j.neuron.2017.02.029.

Howe, M. W., & Dombeck, D. A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, *535*, 505–510, Available at: http://dx.doi.org/10.1038/nature18942.

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M., & Graybiel, A. M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, *500*, 575–579, Available at: http://www.ncbi.nlm.nih.gov/pubmed/23913271. (Accessed 16 September 2013).

Huk, A. C., & Hart, E. (2019). Parsing signal and noise in the brain. *Science*, *364*, 236–237, Available at: http://www.ncbi.nlm.nih.gov/pubmed/31000652.

Jacob, S. N., Ott, T., & Nieder, A. (2013). Dopamine regulates two classes of primate prefrontal neurons that represent sensory signals. *Journal of Neuroscience*, *33*, 13724–13734, Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0210-13.2013.

Jin, X., & Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*, 457–462, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3477867&tool=pmcentrez&rendertype=abstract. (Accessed 27 September 2013).

Jo, Y. S., Heymann, G., & Zweifel, L. S. (2018). Dopamine neurons reflect the uncertainty in fear generalization. *Neuron*, *100*, 916–925.e3. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0896627318308304.

Kato, A., & Morita, K. (2016). Forgetting in reinforcement learning links sustained dopamine signals to motivation. *PLOS Computational Biology*, *12*, 1–41, Available at: http://dx.doi.org/10.1371/journal.pcbi.1005145.

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology*, *40*, 454–462, Available at: http://dx.doi.org/10.1038/npp.2014.193.

Kim, H. F., Ghazizadeh, A., & Hikosaka, O. (2015). Dopamine neurons encoding long-term memory of object value for habitual behavior. *Cell*, *163*, 1165–1175, Available at: http://dx.doi.org/10.1016/j.cell.2015.10.063.

Lak, A., Stauffer, W. R., & Schultz, W. (2016). Dopamine neurons learn relative chosen value from probabilistic rewards. *Elife*, *5*, 1–19.

Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., & Roeper, J. (2008). Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron*, *57*, 760–773, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18341995. (Accessed 17 September 2013).

Lammel, S., Ion, D. I., Roeper, J., & Malenka, R. C. (2011). Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron*, *70*, 855–862, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3112473&tool=pmcentrez&rendertype=abstract. (Accessed 20 September 2013).

Lau, B., Monteiro, T., & Paton, J. J. (2017). The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. *Current Opinion in Neurobiology*, *46*, 241–247, Available at: http://dx.doi.org/10.1016/j.conb.2017.08.015.

Lee, H., Ghim, J.-W., Kim, H., Lee, D., & Jung, M. (2012). Hippocampal neural correlates for values of experienced events. *Journal of Neuroscience*, *32*, 15053–15065.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*, 451–463, Available at: http://dx.doi.org/10.1016/j.neuron.2016.12.040.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*, 1250–1252, Available at: http://dx.doi.org/10.1038/nn.2904. (Accessed 29 September 2013).

Likhtik, E., & Johansen, J. P. (2019). Neuromodulation in circuits of aversive emotional learning. *Nature Neuroscience*, *22*, 1586–1597, Available at: http://dx.doi.org/10.1038/s41593-019-0503-3.

Liu, C., & Kaeser, P. S. (2019). Mechanisms and regulation of dopamine release. *Current Opinion in Neurobiology*, *57*, 46–53, Available at: https://doi.org/10.1016/j.conb.2019.01.001.

Lloyd, K., Becker, N., Jones, M. W., & Bogacz, R. (2012). Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. *Frontiers in Computational Neuroscience*, *6*, 1–10, Available at: http://journal.frontiersin.org/article/10.3389/fncom.2012.00087/abstract.

Lloyd, K., & Dayan, P. (2015). Tamping ramping: Algorithmic, implementational, and computational explanations of phasic dopamine signals in the accumbens. *PLOS Computational Biology*, *11*, 1–34.

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *Elife*, *6*, 1–26.

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J. M. T., Vinson, L. T., Patriarchi, T., et al. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature*, *570*, 65–70, Available at: http://dx.doi.org/10.1038/s41586-019-1235-y.

Morita, K., & Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. *Frontiers in Neural Circuits*, *8*, 36, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3988379&tool=pmcentrez&rendertype=abstract.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*, 1040–1046.

Nasser, H. M., Calu, D. J., Schoenbaum, G., & Sharpe, M. J. (2017). The dopamine prediction error: Contributions to associative models of reward learning. *Frontiers in Psychology*, *8*, 1–17.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., et al. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*, 8145–8157, Available at: http://www.ncbi.nlm.nih.gov/pubmed/26019331. (Accessed 26 October 2017).

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyl, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*, 6235–6242, Available at: http://www.ncbi.nlm.nih.gov/pubmed/15987953. (Accessed 17 September 2013).

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyl, B. I. (2008). Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *Journal of Neuroscience*, *28*, 9619–9631, Available at: http://www.ncbi.nlm.nih.gov/pubmed/18815248. (Accessed 26 September 2013).

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552, Available at: https://pdfs.semanticscholar.org/bfde/4e5ddaa6a968df9984b557baccd7cb38fb82.pdf. (Accessed 15 July 2018).

Pignatelli, M., & Bonci, A. (2015). Role of dopamine neurons in reward and aversion: A synaptic plasticity perspective. *Neuron*, *86*, 1145–1157, Available at: https://www.sciencedirect.com/science/article/pii/S0896627315003657. (Accessed 15 July 2018).

Salinas-Hernández, X. I., Vogel, P., Betz, S., Kalisch, R., Sigurdsson, T., & Duvarci, S. (2018). Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. *Elife*, *7*, 1–25, Available at: https://elifesciences.org/articles/38818.

Schultz, W. (2010). Dopamine signals for reward value and risk: Basic and recent data. *Behavioral and Brain Functions*, *6*, 1–9, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2876988&tool=pmcentrez&rendertype=abstract. (Accessed 1 October 2013).

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*.

Schultz, W., Dayan, P., & Montague, R. (1997). A neural substrate of prediction and reward. *Science (80-)*, *275*, 1593–1599, Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.275.5306.1593 Accessed 19 September 2013.

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., et al. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*, 735–742, Available at: http://www.nature.com/doifinder/10.1038/nn.4538.

Sheafor, P. J., & Gormezano, I. (1972). Conditioning the rabbit's (*Oryctolagus cuniculus*) jaw-movement response: US magnitude effects on URs, CRs, and pseudo-CRs. *Journal of Comparative and Physiological Psychology*, *81*, 449–456, Available at: http://www.ncbi.nlm.nih.gov/pubmed/4649185. (Accessed 9 January 2016.

Sheafor, P. J. (1975). "Pseudoconditioned" jaw movements of the rabbit reflect associations conditioned to contextual background cues. *Journal of Experimental Psychology Animal Behavior Processes*, *1*, 245–260, Available at: http://www.ncbi.nlm.nih.gov/pubmed/1185111. (Accessed 8 January 2016).

Shin, J. H., Kim, D., & Jung, M. W. (2018). Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. *Nature Communications*, *9*.

Skinner, B. (1948). Superstition in the pigeon. *Journal of Experimental Psychology*, *38*, 168–172, Available at: http://psycnet.apa.org/journals/xge/38/2/168/.

Smith, K. S., & Graybiel, A. M. (2016). Habit formation. *Dialogues in Clinical Neuroscience*, *18*, 33–43.

Stark, H., Rothe, T., Wagner, T., & Scheich, H. (2004). Learning a new behavioral strategy in the shuttle-box increases prefrontal dopamine. *Neuroscience*, *126*, 21–29, Available at: http://www.ncbi.nlm.nih.gov/pubmed/15145070. (Accessed 1 October 2013).

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, *20*, 581–589, Available at:http://www.nature.com/doifinder/10.1038/nn.4520.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*, 966–973, Available at: http://www.ncbi.nlm.nih.gov/pubmed/23708143 (Accessed 19 September 2013).

Stuber, G. D., Klanker, M., De Ridder, B., Bowers, M. S., Joosten, R. N., Feenstra, M. G., et al. (2008). Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. *Science (80-)*, *321*, 1690–1692, Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2613864&tool=pmcentrez&rendertype=abstract. (Accessed 1 October 2013).

Sugam, J. A., Day, J. J., Wightman, R. M., & Carelli, R. M. (2012). Phasic nucleus accumbens dopamine encodes risk-based decision-making behavior. *Biological Psychiatry*, *71*, 199–205, Available at: http://dx.doi.org/10.1016/j.biopsych.2011.09.029.

Syed, E. C. J. J., Grima, L. L., Magill, P. J., Bogacz, R., Brown, P., & Walton, M. E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nature Neuroscience*, *19*, 34–36, Available at:http://www.nature.com/doifinder/10.1038/nn.4187.

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, *95*, 1395–1405.e3. Available at: https://doi.org/10.1016/j.neuron.2017.08.025.

Tian, J., Huang, R., Cohen, J. Y., Osakada, F., Kobak, D., Machens, C. K., et al. (2016). Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron*, *91*, 1374–1389, Available at: http://dx.doi.org/10.1016/j.neuron.2016.08.018.

Todd, M. T., Niv, Y., & Cohen, J. D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In *Advances in neural information processing systems (vol. 21)* (pp. 1689–1696). Available at: https://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observable-environments-through-dopaminergic-reinforcement. (Accessed 25 October 2017).

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron*, *89*, 695–710.

Westbrook, A., & Frank, M. (2018). Dopamine and proximity in motivation and cognitive control. *Current Opinion in Behavioral Sciences*, *22*, 28–34, Available at: https://doi.org/10.1016/j.cobeha.2017.12.011.

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. *Journal of Neuroscience*, *27*, 8181–8183, Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1671-07.2007. (Accessed 15 July 2018).

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*, 464–476, Available at: http://www.ncbi.nlm.nih.gov/pubmed/16715055. (Accessed 15 July 2018).

Yin, H. H., Zhuang, X., & Balleine, B. W. (2006). Instrumental learning in hyperdopaminergic mice. *Neurobiology of Learning and Memory*, *85*, 283–288, Available at: http://www.ncbi.nlm.nih.gov/pubmed/16423542. (Accessed 17 September 2013).