*Title:*

Reinforcement learning of state representation and value: the power of random feedback and biological constraints

*Authors:*

Takayuki Tsurumi[1,2], Ayaka Kato[2,3], Arvind Kumar[2,4], & Kenji Morita[1,2,5*]

*Affiliations:*

[1] Physical and Health Education, Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2] Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology, Tancha, Okinawa, Japan

[3] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[4] Division of Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

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

[*]*Corresponding author:*

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

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

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

morita@p.u-tokyo.ac.jp

*Author contributions*

Conceptualization: KM; Formal analysis: KM, TT; Investigation: KM, TT, AyK; Writing – original draft: KM; Writing – review & editing: KM, TT, AyK, ArK

*Acknowledgements*

The authors thank Dr. Kenji Doya for valuable suggestions. KM was supported by Grants-in-Aid for Scientific Research 23H03295 from Japan Society for the Promotion of Science (JSPS) and the Naito Foundation. AyK was supported by JSPS Overseas Research Fellowships. ArK was partially funded by Digital Futures (KTH) grant and StratNeuro SRA.

# Abstract

How external/internal 'state' is represented in the brain is crucial, since appropriate representation enables goal-directed behavior. Recent studies suggest that state representation and state value can be simultaneously learnt through reinforcement learning (RL) using reward-prediction-error in recurrent-neural-network (RNN) and its downstream weights. However, how such learning can be neurally implemented remains unclear because training of RNN through the 'backpropagation' method requires downstream weights, which are biologically unavailable at the upstream RNN. Here we show that training of RNN using random feedback instead of the downstream weights still works because of the 'feedback alignment', which was originally demonstrated for supervised learning. We further show that if the downstream weights and the random feedback are biologically constrained to be non-negative, learning still occurs without feedback alignment because the non-negative constraint ensures loose alignment. These results suggest neural mechanisms for RL of state representation/value and the power of random feedback and biological constraints.

# Introduction

Multiple lines of studies have suggested that Temporal-Difference-Reinforcement-Leaning (TDRL) is implemented in the cortico-basal ganglia-dopamine(DA) circuits in the way that DA represents TD reward-prediction-error (RPE) [1-5] and DA-dependent plasticity of cortico-striatal synapses represents TD-RPE-dependent update of state/action values [6-8]. Traditionally, TDRL in the cortico-basal ganglia-DA circuits was considered to serve only for relatively simple behavior. However, subsequent studies suggested that more sophisticated, apparently goal-directed/model-based behavior can also be achieved by TDRL if states are appropriately represented [9-11] and that DA signals indeed reflect model-based predictions [12]. Conversely, state representation-related issues could potentially cause behavioral or mental-health problems [13-17]. Early modeling studies treated state representations appropriate to the situation/task as given ('handcrafted' by the authors), but representation itself should be learnt in the brain [18]. Recently it was shown that appropriate state representation can be learnt through RL in a recurrent neural network (RNN) by minimization of squared value-error without explicit teacher/target [11, 19], while state value can be simultaneously learnt in the downstream of RNN.

However, whether such a learning method, named the value-RNN [19], can be implemented in the brain remains unclear, because there are problems in terms of biological plausibility. A major problem, among others, is that the update rule proposed in the previous work for the connections onto the 'neurons' in the RNN [19], derived from the gradient-descent error-'backpropagation' (hereafter referred to as backprop) method [20, 21], involves the weights of the connections from these RNN units onto the downstream value-encoding unit. Given that the state-representing RNN and the value-encoding unit are implemented by the intra-cortical circuit and the striatal neurons, respectively, as generally suggested [1, 22, 23], this means that the update (plasticity) rule for intra-cortical connections involves the downstream cortico-striatal synaptic strengths, which would not be able to be accessed from the cortex. Indeed, this is an example of the long-standing difficulty in biological implementation of backprop [24, 25], in which update of upstream connections requires biologically unavailable downstream connection strengths.

Recently, a potential solution for this difficulty has been proposed [26] (see also [27-29] for other potential solutions). Specifically, in supervised learning of feed-forward network, it was shown that when the downstream connection strengths used for updating upstream connections in backprop were replaced with fixed random strengths, comparable learning performance was still achieved [26]. This was suggested to be because the information of the introduced fixed random strengths transferred, through learning, to the upstream connections and then to the downstream feed-forward connections so that these feed-forward connections became aligned to the random feedback strengths, and thus in turn, the random feedback can play the same role as the one played by the downstream connection strengths in backprop. This mechanism was named the 'feedback alignment' [26], and was subsequently shown to

work also in supervised learning of RNN [30] and proposed to be neurally implemented [31] (in a different way from the present study as we discuss in the Discussion).

The value-RNN [11, 19], the above-introduced simultaneous RL of state values and state representation through minimization of squared value-error, differs from supervised learning considered in these previous feedback-alignment studies in two ways: i) it is TD learning, i.e., it approximates the true error by the TD-RPE because the true error, or true state value, is unknown, and ii) it uses a scalar error (TD-RPE) rather than a vector error. Therefore it was nontrivial whether the feedback alignment mechanism could work also for the value-RNN. In the present work, we first examined this, demonstrating that it does work and providing a mechanistic insight into how it works.

After that, we further addressed other biological-plausibility problems. Specifically, we imposed biological constraints that the downstream (cortico-striatal) weights and the fixed random feedback, as well as the activities of neurons in the RNN, were all non-negative. Moreover, we also remedied the non-monotonic dependence of the update of RNN connection-strength on post-synaptic neural activity. We then found, unexpectedly, that the non-negative constraint appeared to aid, rather than degrade, the learning by ensuring that the downstream weights and the fixed random feedback are loosely aligned even without operation of the feedback alignment mechanism. These results suggest how learning of state representation and value can be neurally implemented, more specifically, through synaptic plasticity depending on DA, which represents TD-RPE, in the cortex and the striatum.

## Results

### *Consideration of the value-RNN with fixed random feedback*

We considered an implementation of the value-RNN in the cortico-basal ganglia circuits (Fig. 1). A cortical region/population is supposed to represent information of sensory observation (***o***) and send it to another cortical region/population, which has rich recurrent connections and therefore can be approximated by an RNN. Activities of neurons in the RNN (***x***) are supposed to learn to represent states, through updates of the strengths of recurrent connections **A** and feed-forward connections **B**. The activity of a population of striatal neurons that receive inputs from the RNN is supposed to learn to represent the state values ($v$), by learning the weights of cortico-striatal connections (from the RNN to the striatal neurons) (***w***) indicating the value weights. DA neurons in the ventral tegmental area (VTA) receive (direct and indirect) inputs from the striatum and other structure conveying information of obtained reward ($r$), and thereby the activity of the DA neurons, as well as released DA, represents TD-RPE ($\delta$). TD-RPE-representing DA is released in the striatum and also in the cortical RNN through mesocorticolimbic projections, and used for modifying the strengths of cortical recurrent and feed-forward connections (**A** and **B**) and cortico-striatal connections (***w***).

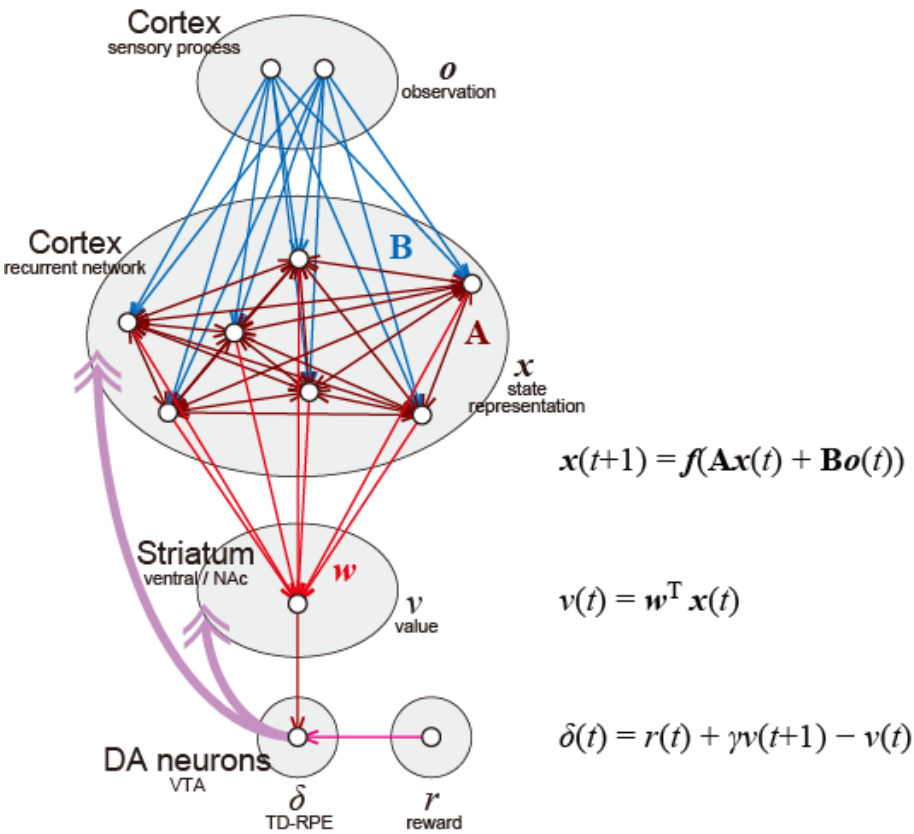


Figure 1

**Figure 1**
Implementation of the value-RNN in the cortico-basal ganglia-DA circuits.

In the original value-RNN [11, 19], update rule for the connections onto the RNN (**A** and **B**) requires the (gradually changing) value weights (***w***), but this is biologically implausible because the corticostriatal synaptic strengths are not available in the cortex as discussed above. Therefore, we considered a modified value-RNN by replacing the cortico-striatal weights used in the updates of intra-cortical connections with fixed random strengths (***c***). Besides, the original value-RNN adopted a learning rule called the backpropagation through time (BPTT) [32], in which the error in the output needs to be incrementally accumulated in the temporally backward order, but such an acausality is also biologically implausible, as previously pointed out [30]. Therefore, we instead used an online learning rule, which considers only the influence of the recurrent weights at the previous time step (see the Methods for details and equations).

### *Simulation of a Pavlovian cue-reward association task with variable inter-trial intervals*

We compared the learning of the modified value-RNN with fixed random feedback (referred to as VRNNrf) and the value-RNN with backprop (referred to as VRNNbp), both of which adopted the online learning rule rather than the BPTT, and also untrained RNN. The number of RNN units was set to 7 for all the cases. Traditional TD-RL agents with punctate state representation (called the complete serial compound, CSC [1, 33]) were also compared. We simulated a Pavlovian cue-reward association task, in which a cue was followed by a reward three time-steps later, and inter-trial interval (i.e., reward to next cue) was randomly chosen from 4, 5, 6, or 7 time-steps (Fig. 2A). In this task, states can be defined by relative timings from the cue, and we estimated the true state values through simulations according to the definition of state value, i.e., expected cumulative discounted future rewards [34] (Fig. 2B, black line). Expected TD-RPE calculated from these estimated true values (Fig. 2B, red line) was almost 0 at any states, as expected. Agent having punctate/CSC state representation and state values without continuation between trials (i.e., the value of the last state in a trial was not updated by TD-RPE upon entering the next trial) developed positive values between cue and reward, and abrupt TD-RPE upon cue (Fig. 2C). Agent having punctate/CSC state representation and continuously updated state values across trials developed positive values also for states in the inter-trial interval (Fig. 2D). VRNNbp developed state values between cue and reward, and to some extent in the inter-trial interval, and showed abrupt TD-RPE upon cue and smaller TD-RPE upon reward (Fig. 2E). This indicates that this agent largely learned the task structure, confirming the previously proposed effectiveness of value-RNN in this different task.

VRNNrf, having fixed random feedback instead of backprop-based feedback, developed state values that were largely similar to, although smaller (on average across simulations) than, those developed by VRNNbp (Fig. 2F black line). VRNNrf generated abrupt TD-RPEs upon cue and reward, again similarly to VRNNbp although the relative size of reward-response was (on average) larger (Fig. 2F red line). As a comparison, agent with untrained RNN developed (on average) even smaller state values and larger relative size of TD-RPE upon reward (Fig. 2G). These results indicate that value-

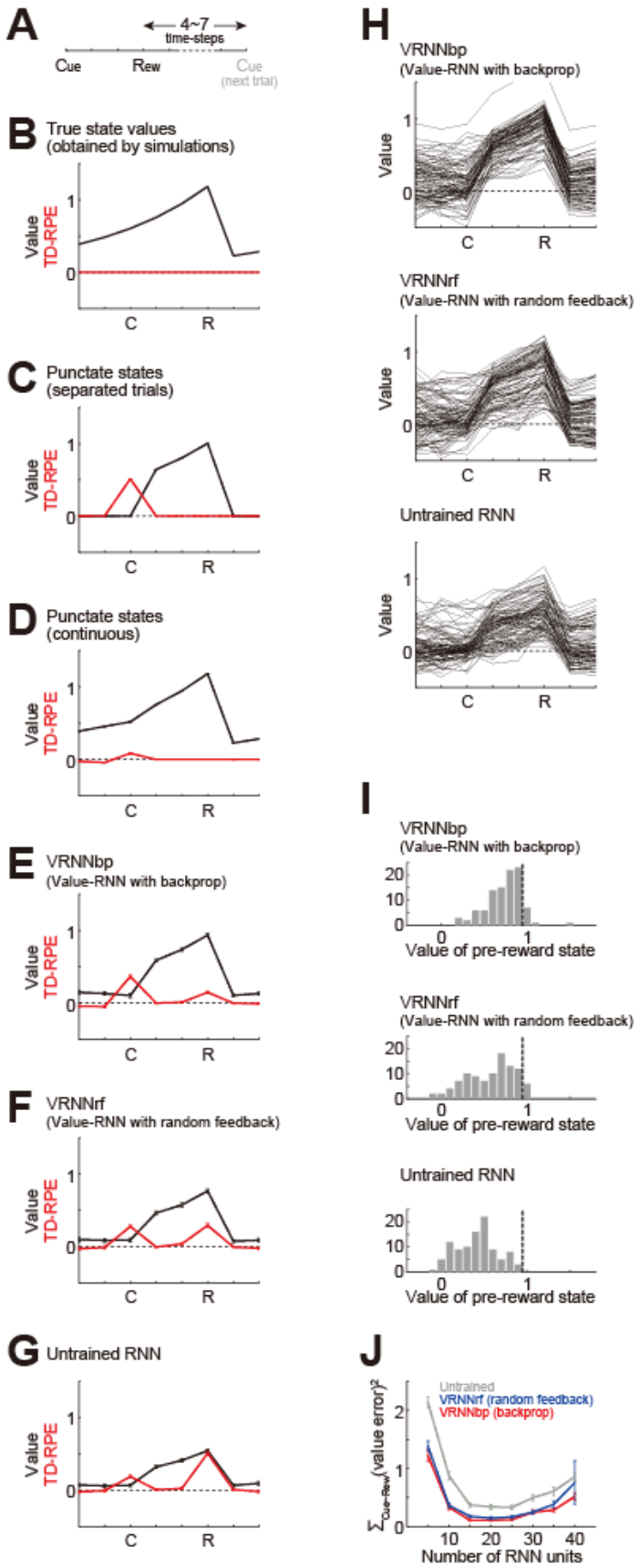


Figure 2

**Figure 2**

Simulation of a Pavlovian cue-reward association task. **(A)** Simulated task with variable inter-trial intervals. **(B)** Black line: Estimated true values of states, defined by relative timings from the cue, through simulations according to the definition of state value, i.e., expected cumulative discounted future rewards. Red line: TD-RPEs calculated from the estimated true state values. **(C-G)** State values (black lines) and TD-RPEs (red lines) at 1000-th trial, averaged across 100 simulations (error-bars indicating ± SEM across simulations), in different types of agent: (C) TD-RL agent having punctate/CSC state representation and state values without continuation between trials (i.e., the value of the last state in a trial was not updated by TD-RPE upon entering the next trial); (D) TD-RL agent having punctate/CSC state representation and continuously updated state values across trials; (E) Value-RNN with backprop (VRNNbp). The number of RNN units was 7 (same applied to (F,G); (F) Value-RNN with fixed random feedback (VRNNrf); (G) Agent with untrained RNN. **(H)** State values at 1000-th trial in individual simulations of VRNNbp (top), VRNNrf (middle), and untrained RNN (bottom). **(I)** Histograms of the value of the pre-reward state (i.e., the state one-time step before the reward state) at 1000-th trial in individual simulations of the three models. The vertical black dashed lines indicate the true value of the pre-reward state (estimated through simulations). **(J)** Learning performance of VRNNbp (red line), VRNNrf (blue line), and the untrained RNN (gray line) when the number of RNN units was varied from 5 to 40 (horizontal axis). Leaning performance was measured by the sum of squares of differences between the state values developed at 1000-th trial by each of these three types of agent and the estimated true state values between cue and reward (vertical axis), averaged over 100 simulations (error-bars indicating ± SEM across simulations).

RNN could be trained by fixed random feedback at least to a certain extent, although somewhat less effectively (as maybe expected) than by backprop-based feedback. Figure 2H shows state values developed in individual simulations of VRNNbp (top), VRNNrf (middle), and untrained RNN (bottom), and Figure 2I shows the histograms of the value of the pre-reward state (i.e., one time-step before the state where reward was obtained) developed in individual simulations of these three models. These figures indicate that VRNNrf did not tend to develop moderately smaller state values than VRNNbp in each simulation. Rather, state values developed in VRNNrf were largely comparable to those developed in VRNNbp once they were successfully learned but the success rate was smaller than VRNNbp while still larger than the untrained RNN.

So far, we examined the cases where the number of RNN units was 7. We compared the learning performance of VRNNbp, VRNNrf, and untrained RNN when the number of RNN units was varied from 5 to 40. Leaning performance was measured by the sum of squares of differences between the state values developed by each of these three types of agents and the estimated true state values (Fig. 2B) between cue and reward. As shown in Fig. 2J, on average across simulations, VRNNbp generally achieved the highest performance, but VRNNrf also exhibited largely comparable performance and it always outperformed the untrained RNN. As the number of RNN units increased from 5 to 15, all these three agents improved their performance, while additional increase of RNN units to 20 or 25 resulted in smaller changes. Further increase of RNN units caused decrease in the mean performance in all the three agents, and when the number of RNN units was increased to 45, there were occasions where learning appeared to diverge. We will discuss these in the Discussion.

### *Occurrence of feedback alignment and an intuitive understanding of its mechanism*

We questioned if feedback alignment underlay the learnability of VRNNrf. Returning to the case with 7 RNN units, we examined whether the value weight vector $\boldsymbol{w}$ became aligned to the random feedback vector $\boldsymbol{c}$ in VRNNrf, by looking at the changes in the angle between these two vectors across trials. As shown in Fig. 3A, this angle, averaged across simulations, decreased over trials, indicating that the value weight $\boldsymbol{w}$ indeed tended to become aligned to the random feedback $\boldsymbol{c}$. We then examined whether better alignment of $\boldsymbol{w}$ to $\boldsymbol{c}$ related to better development of state value by looking at the relation between the angle between $\boldsymbol{w}$ and $\boldsymbol{c}$ and the value of the pre-reward state at 1000-th trial. As shown in Fig. 3B, there was a negative correlation such that the smaller the angle was (i.e., more aligned), the larger the state value tended to be ($r = -0.288$, $p = 0.00362$), in line with our expectation. These results indicate that the mechanism of feedback alignment, previously shown to work for supervised learning, also worked for TD learning of value weights and recurrent/feed-forward connections.

How did the feedback alignment mechanistically occur? We made an attempt to obtain an intuitive understanding. Assume that positive TD-RPE ($\delta(t) > 0$) is generated at a state, $S$ ($= \boldsymbol{x}(t)$), in a task trial. Because of the update rule for $\boldsymbol{w}$ ($\boldsymbol{w} \leftarrow \boldsymbol{w} + a\delta(t)\boldsymbol{x}(t)$), $\boldsymbol{w}$ is updated in the direction of $\boldsymbol{x}(t)$. Next, what

is the effect of updates of recurrent/feed-forward connections (**A** and **B**) on $\boldsymbol{x}$? For simplicity, here we consider the case where observation is null ($\boldsymbol{o} = \mathbf{0}$) and so $\boldsymbol{x}(t) = \boldsymbol{f}(\mathbf{A}\boldsymbol{x}(t-1))$ holds (but similar argument can be done in the case where observation is not null). If **A** is replaced with its updated one, it can be calculated that $i$-th element of $\mathbf{A}\boldsymbol{x}(t-1)$ will hypothetically change by $c_i \times$ (a positive value) (technical note: the value is $a\delta(t)\{\Sigma_j x_j(t-1)^2\}(0.5 + x_i(t))(0.5 - x_i(t))$ which is positive unless $x(t-1) = \mathbf{0}$), and therefore the vector $\mathbf{A}\boldsymbol{x}(t-1)$ as a whole will hypothetically change by a vector that is in a relatively close angle with $\boldsymbol{c}$ (in a sense that, for example, $[c_1\ c_2\ c_3]^T$ and $[0.5c_1\ 1.2c_2\ 0.8c_3]^T$ are in a relatively close angle in the same quadrant). Then, because $f$ is a monotonically increasing sigmoidal function, $\boldsymbol{x}(t) = \boldsymbol{f}(\mathbf{A}\boldsymbol{x}(t-1))$ will also hypothetically change by a vector that is in a relatively close angle with $\boldsymbol{c}$. This was indeed the case in our simulations as shown in Fig. 3C.

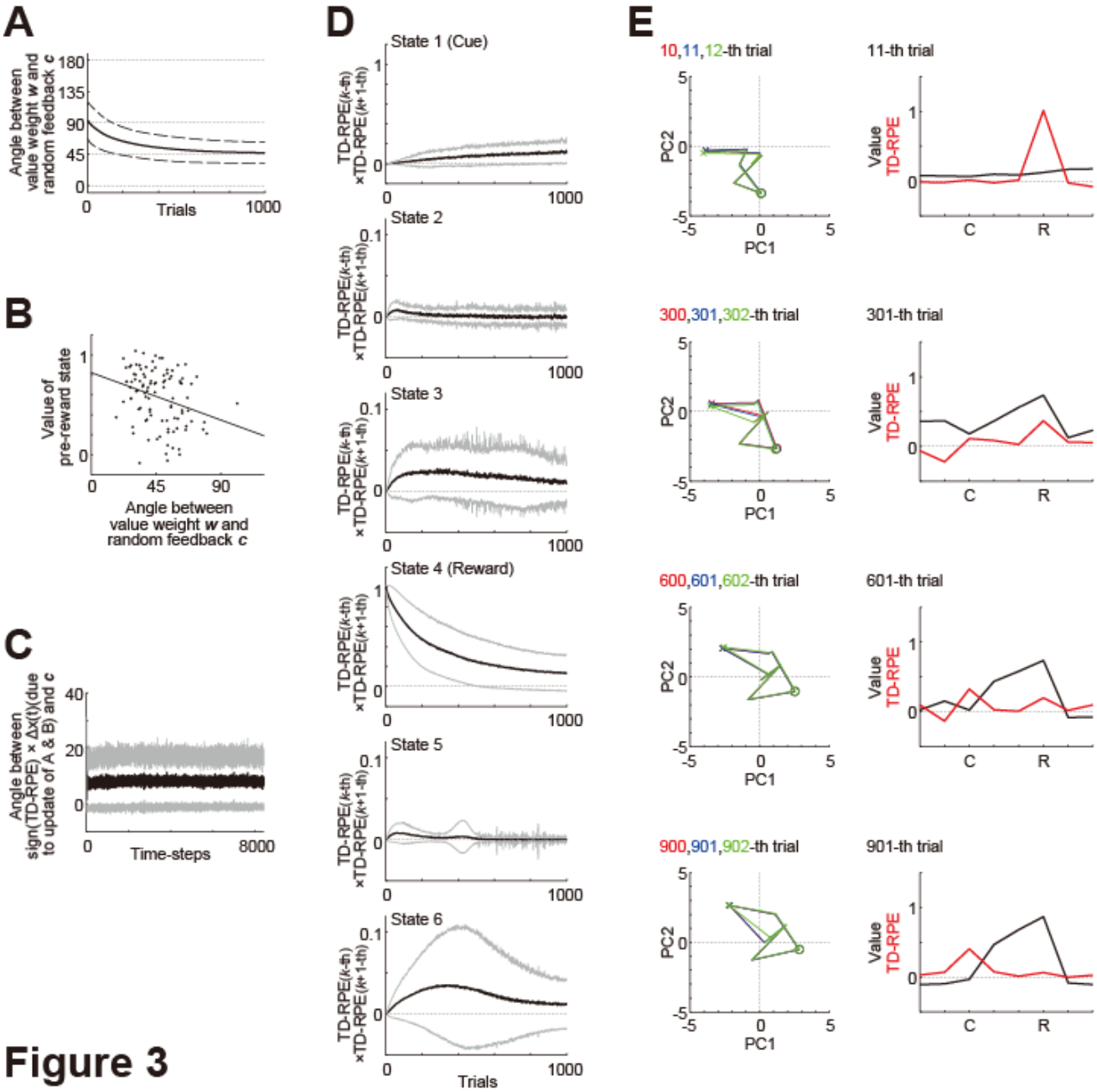


**Figure 3**

**Figure 3**
Occurrence of feedback alignment and an intuitive understanding of its mechanism. **(A)** Over-trial changes in the angle between the value-weight vector $\boldsymbol{w}$ and the fixed random feedback vector $\boldsymbol{c}$ in the simulations of VRNNrf (7 RNN units). The solid line and the dashed lines indicate the mean and ± SD across 100 simulations, respectively. **(B)** The relation between the angle between $\boldsymbol{w}$ and $\boldsymbol{c}$ (horizontal axis) and the value of the pre-reward state (vertical axis) at 1000-th trial. The dots indicate the results of individual simulations, and the line indicates the regression line. **(C)** Angle between the hypothetical change in $\boldsymbol{x}(t) = \boldsymbol{f}(\mathbf{A}\boldsymbol{x}(t-1),\mathbf{B}\boldsymbol{o}(t-1))$ in case **A** and **B** were replaced with their updated ones, multiplied with the sign of TD-RPE (sign($\delta(t)$)), and the fixed random feedback vector $\boldsymbol{c}$ across time-steps. The black thick line and the gray lines indicate the mean and ± SD across 100 simulations, respectively (same applied to (D)). **(D)** Multiplication of TD-RPEs in successive trials at individual states (top: cue, 4th from the top: reward). Positive or negative value indicates that TD-RPEs in successive trials have the same or different signs, respectively. **(E)** *Left*: RNN trajectories mapped onto the primary and secondary principal components (horizontal and vertical axes, respectively) in three successive trials (red, blue, and green lines (heavily overlapped)) at different phases in an example simulation (10th-12th, 300th-302th, 600th-602th, and 900th-902th trials from top to bottom). The crosses and circles indicate the cue and reward states, respectively. *Right*: State values (black lines) and TD-RPEs (red lines) at 11th, 301th, 601th, and 901th trial.

In this way, at state $S$ where TD-RPE is positive, $\boldsymbol{w}$ is updated in the direction of $\boldsymbol{x}(t)$, and $\boldsymbol{x}(t)$ will hypothetically change by a vector that is in a relatively close angle with $\boldsymbol{c}$ if **A** is replaced with its updated one. Then, if the update of $\boldsymbol{w}$ and the hypothetical change in $\boldsymbol{x}(t)$ due to the update of **A** could be integrated, $\boldsymbol{w}$ would become aligned to $\boldsymbol{c}$ (if TD-RPE is instead negative, $\boldsymbol{w}$ is updated in the opposite direction of $\boldsymbol{x}(t)$, and $\boldsymbol{x}(t)$ will hypothetically change by a vector that is in a relatively close angle with $-\boldsymbol{c}$, and so the same story holds in the end).

There is, however, a caveat regarding how the update of $\boldsymbol{w}$ and the hypothetical change in $\boldsymbol{x}(t)$ can be integrated. Although technical, here we briefly describe it, and a possible solution. The updates of $\boldsymbol{w}$ and **A** use TD-RPE, which is calculated based on $v(t) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(t)$ and $v(t+1) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(t+1)$, and so $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t+1)$ should already be determined beforehand. Therefore, the hypothetical change in $\boldsymbol{x}(t)$ due to the update of **A**, described in the above, does not actually occur (this was why we mentioned 'hypothetical') and thus cannot be integrated with the update of $\boldsymbol{w}$. Nevertheless, integration could still occur across successive trials, at least to a certain extent. Specifically, although TD-RPEs at $S$ in successive trials would generally differ from each other, they would still tend to have the same sign, as was indeed the case in our simulations (Fig. 3D). Also, although the trajectories of RNN activity ($\boldsymbol{x}$) in successive trials would differ, we could expect a certain level of similarity because the RNN is entrained by observation-representing inputs, again as was indeed the case in our example simulation (Fig. 3E). Then, the hypothetical change in $\boldsymbol{x}(t)$ due to the update of **A**, considered above, could become a reality in the next trial, to a certain extent, and could thus be integrated into the update of $\boldsymbol{w}$, explaining the occurrence of feedback alignment.

***Simulation of tasks with probabilistic structures of reward timing/existence***

We also simulated two tasks (Fig. 4A) that were qualitatively similar to (though simpler than) the two tasks examined in previous experiments [35] and modeled by the original value-RNN with backprop [19]. In our task 1, a cue was always followed by a reward either two or four time-steps later with equal probabilities. Task 2 was the same as task 1 except that reward was omitted with 40% probability. In task 1, if reward was not given at the early timing (i.e., two-steps later than cue), agent could predict that reward should be given at the late timing (i.e., four-steps later than cue), and thus TD-RPE upon reward at the late timing is expected to be smaller than TD-RPE upon reward at the early timing (if agent perfectly learned the task structure, TD-RPE upon reward at the late timing should be 0). By contrast, in task 2, if reward was not given at the early timing, it might indicate that reward was given at the late timing but might instead indicate that reward was omitted in that trial, and thus TD-RPE upon reward at the late timing is expected to exist and can even be larger than TD-RPE upon reward at the early timing.

In these tasks, states can be defined in the following way. There were two types of trials, with early or late reward, in task 1, and additionally one more type of trial, without reward, in task 2 (Fig. 4Ba, top). For each timing in each of these trial types, its value, i.e., expected discounted cumulative future rewards, can be estimated through simulations (Fig. 4Ba, bottom). Agent could not know the current trial type until receiving reward at the early timing or the late timing or receiving no reward at both timings. Until these timings, agent could have probabilistic belief about the current trial type, e.g., 50% in the trial with early reward and 50% in the trial with late reward (in task 1) or 30% in the trial with early reward, 30% in the trial with late reward, and 40% in the trial without reward (in task 2) (Fig. 4Bb). States can be defined by incorporating these probabilistic beliefs at each timing (Fig. 4Bc, top), and state values (Fig. 4Bc, bottom: expected discounted cumulative future rewards, estimated through simulations) should theoretically match an integration (multiplication) of the values of each trial type (Fig. 4Ba, bottom) with the probabilistic beliefs (Fig. 4Bb). Expected TD-RPE calculated from these estimated state values (Fig. 4C) exhibited features that matched the conjecture mentioned above: in task 1, TD-RPE upon reception of late reward, which was actually 0, was smaller than TD-RPE upon reception of early reward, whereas in task 2, TD-RPE upon reception of late reward was larger than TD-RPE upon reception of early reward.

The previous experimental work [35] has shown that VTA DA neurons exhibited similar activity patterns to the abovementioned TD-RPE patterns, and the theoretical work [19] has shown that the original value-RNN with backprop could reproduce such TD-RPE patterns. We examined what TD-RPE patterns were generated in the agents with punctate/CSC representation, VRNNbp, VRNNrf, and untrained RNN (12 RNN units in all cases) in our simulated two tasks. VRNNbp developed similar TD-RPE patterns (smaller TD-RPE upon late than early timing in task 1 but opposite pattern in task 2) (Fig. 4F), qualitatively reproducing the result of the previous work [19]. Crucially, VRNNrf also developed similar TD-RPE patterns (Fig. 4G), indicating that this agent with random feedback could

also learn the distinct structures of the two tasks. By contrast, agents with punctate/CSC state representation without or with continuous value update across trials (Fig. 4D,E), as well as agent with untrained RNN (Fig. 4H), could not develop such patterns well.

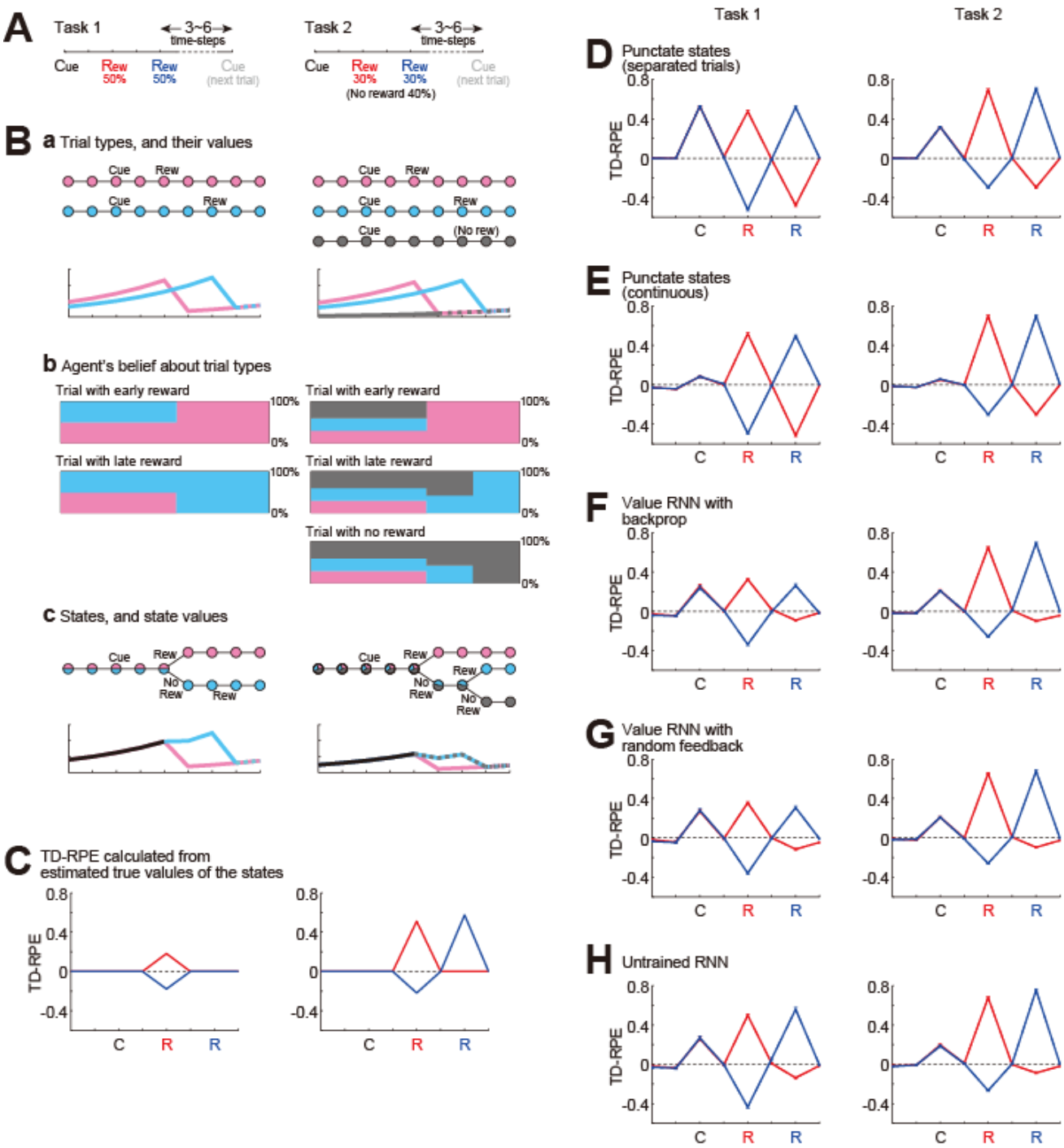


**Figure 4**

**Figure 4**
Simulation of two tasks having probabilistic structures, which were qualitatively similar to the two tasks examined in experiments [35] and modeled by value-RNN [19]. **(A)** Simulated two tasks, in which reward was given at the early or the late timing with equal probabilities in all the trials (task 1) or 60% of trials (task 2). **(B) (a)** *Top*: Trial types. Two trial types (with early reward and with late reward) in task 1 and three trial types (with early reward, with late reward, and without reward) in task 2. *Bottom*: Value (expected discounted cumulative future rewards) of each timing in each trial type. **(b)** Agent's probabilistic belief about the current trial type, in the case where agent was in fact in the trial with early reward (top row), the trial with late reward (second row), or the trial without reward (third row in task 2). **(c)** *Top*: States defined by considering the probabilistic beliefs at each timing from cue. *Bottom*: State values (expected discounted cumulative future rewards, estimated through simulations), which should theoretically match an integration (multiplication) of the values of each trial type (shown in (a)-bottom) with the probabilistic beliefs (shown in (b)). **(C)** Expected TD-RPE calculated from the estimated true values of the states for task 1 (left) and task 2 (right). Red lines: case where reward was given at the early timing, blue lines: case where reward was given at the late timing. **(D-H)** TD-RPEs at the latest trial within 1000 trials in which reward was given at the early timing (red lines) or the late timing (blue lines), averaged across 100 simulations (error-bars indicating ± SEM across simulations), in the different types of agent: (D,E) TD-RL agent having punctate/CSC state representation and state values without (D) or with (E) continuation between trials; (F) VRNNbp. The number of RNN units was 12 (same applied to (G,H); (G) VRNNrf; (H) Untrained RNN.

### *Value-RNN with further biological constraints*

So far, the activities of neurons in the RNN ($\boldsymbol{x}$) were initialized to pseudo standard normal random numbers, and thereafter took numbers in the range between −0.5 and 0.5 that was the range of the sigmoidal input-output function. The value weights ($\boldsymbol{w}$) could also take both positive and negative values since no constraint was imposed. The fixed random feedback in VRNNrf ($\boldsymbol{c}$) was generated by pseudo standard normal random numbers, and so could also be positive or negative. Negativity of the neurons' activities and the value weights could potentially be regarded as inhibitory or smaller-than-baseline quantities. However, because neuronal firing rate is non-negative and cortico-striatal projections are excitatory, it would be biologically more plausible to assume that the activities of neurons in the RNN and the value weights are non-negative. As for the fixed random feedback, if it is negative, the update rule becomes anti-Hebbian under positive TD-RPE, and so assuming non-negativity would be plausible since Hebbian property has been suggested for rapid plasticity of cortical synapses [36]. There was another issue in the update rule for recurrent and feed-forward connections, derived from the gradient descent. Specifically, the dependence on the post-synaptic activity was non-monotonic, maximized at the middle of the range of activity. It would be more biologically plausible to assume monotonic dependence.

In order to address these issues, we considered revised models. We first considered a revised VRNNbp, referred to as revVRNNbp, in which the RNN activities and the value weights were constrained to be non-negative, while the non-monotonic dependence of the update rule on the post-

synaptic activity remained unchanged (Fig. 5A). We then considered a revised VRNNrf, referred to as bioVRNNrf, in which the fixed random feedback, as well as the RNN activities and the value weights, were constrained to be non-negative, and also the update rule was modified so that the dependence on the post-synaptic activity became monotonic (with saturation) (Fig. 5B).

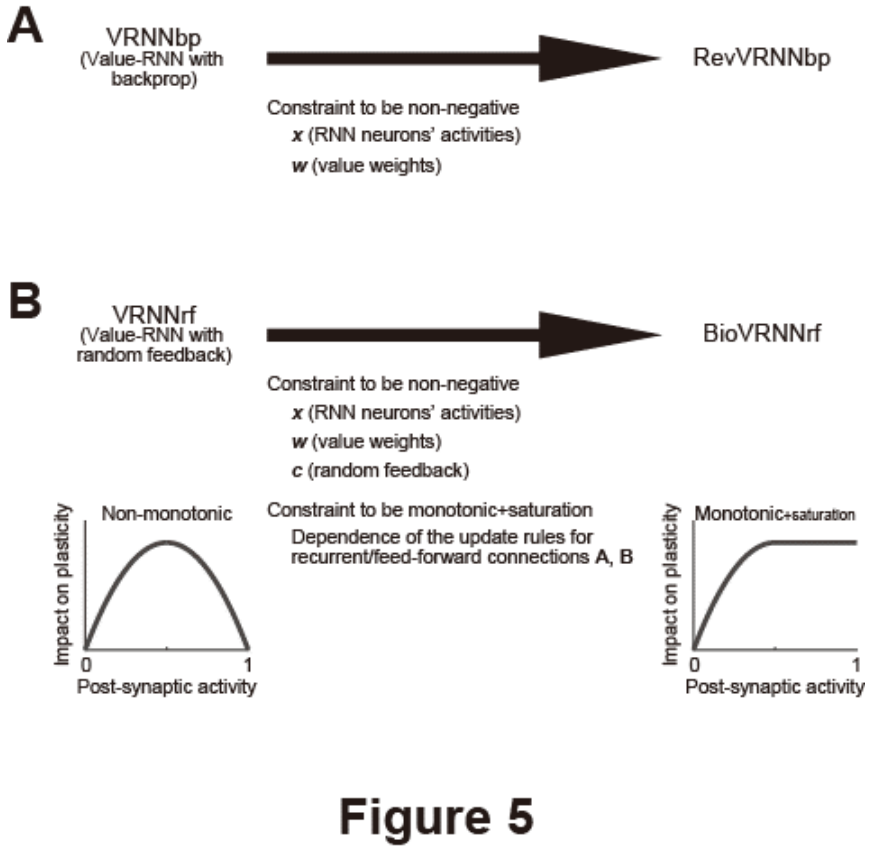


**Figure 5**
Modified value-RNN models with further biological constraints. **(A)** RevVRNNbp: VRNNbp (value-RNN with backprop) was modified so that the activities of neurons in the RNN (***x***) and the value weights (***w***) became non-negative. **(B)** BioVRNNrf: VRNNrf (value-RNN with fixed random feedback) was modified so that ***x*** and ***w***, as well as the fixed random feedback (***c***), became non-negative and also the dependence of the update rules for recurrent/feed-forward connections (**A** and **B**) on post-synaptic activity became monotonic + saturation.

We examined how these revised models, in comparison with untrained RNN that also had non-negative constraint for ***x*** and ***w***, performed in the Pavlovian cue-reward association task examined above (the number of neurons/trials were set to 12/1500). RevVRNNbp well developed state values toward reward (Fig. 6A). BioVRNNrf also developed state values to a largely comparable extent (Fig. 6B). By contrast, untrained RNN could not develop such a pattern of state values (Fig. 6C). This, however, could be because initially set recurrent/feed-forward connections were far from those learned in the value-RNNs. Therefore, as a more strict control, we conducted simulations of untrained RNN with non-negative ***x*** and ***w***, where in each simulation the recurrent/feed-forward connections were set to be those shuffled from the learnt connections in a simulation of bioVRNNrf. Untrained RNN with this setting performed somewhat better than the original untrained RNN case (Fig. 6D), but still worse than revVRNNbp and bioVRNNrf. We varied the number of neurons in the RNN, and compared the performance (sum of squared errors from the true state values) of revVRNNbp and bioVRNNrf, in comparison with untrained RNN (both naive one and the one with shuffled learnt connections from bioVRNNrf). As shown in Fig. 6E, regardless of the number of neurons, the performance of bioVRNNrf was largely comparable to that of revVRNNbp, and better than the performance of untrained RNN of both kinds. Figure 6F shows the mean of the elements of the recurrent and feed-

forward connections at 1500-th trial in the different models. As shown in the figure, these connections (initialized to pseudo standard normal random numbers) were learnt to become negative on average, in revVRNNbp and more prominently in bioVRNNrf. This learnt negative-dominance (inhibition-dominance) could possibly be related, e.g., through prevention of excessive activity, to the good performance of bioVRNNrf and also the better performance of the untrained RNN with connections shuffled from bioVRNNrf than that of the naive untrained RNN.

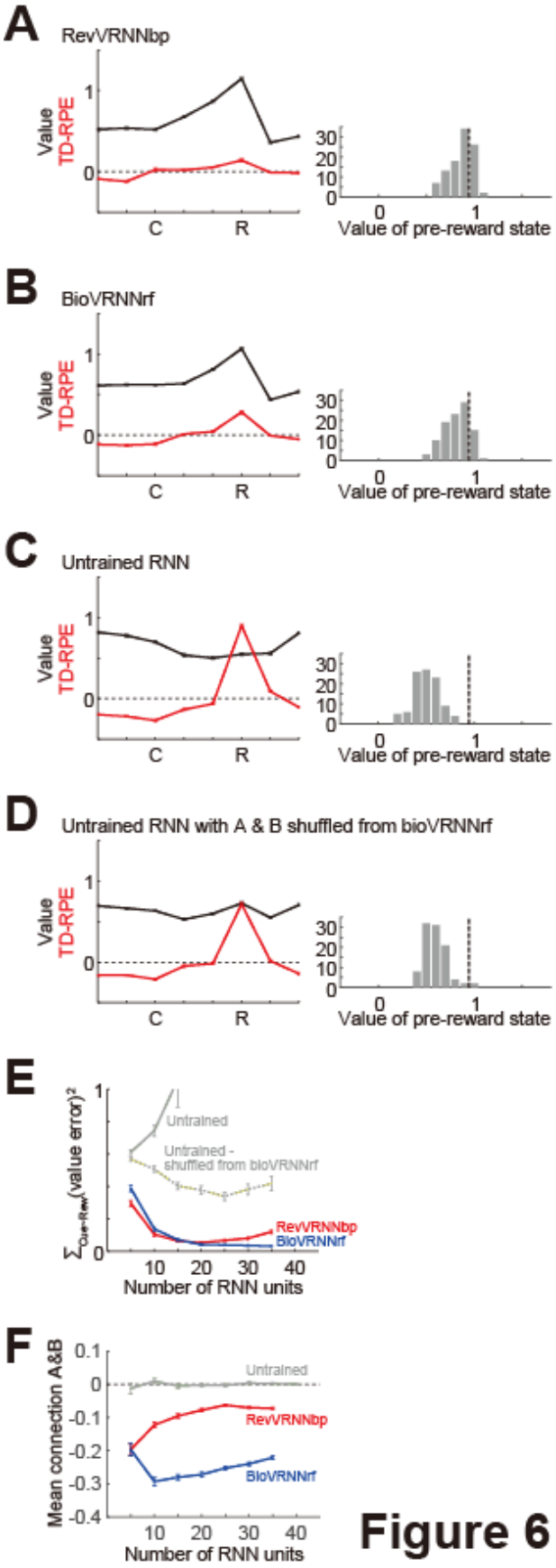


Figure 6

**Figure 6**
Performances of the modified value-RNN models in the cue-reward association task, in comparison with untrained RNN that also had the non-negative constraint. **(A-D)** State values (black lines) and TD-RPEs (red lines) at 1500-th trial in revVRNNbp (A), bioVRNNrf (B), untrained RNN with ***x*** and ***w*** constrained to be non-negative (C), and untrained RNN with non-negative ***x*** and ***w*** and having connections shuffled from those learnt in bioVRNNrf (D). The number of RNN units was 12 in all the cases. Error-bars indicate mean ± SEM across 100 simulations (same applied to (E,F)). The right histograms show the across-simulation distribution of the value of the pre-reward state in each model. The vertical black dashed lines in the histograms indicate the true value of the pre-reward state (estimated through simulations). **(E)** Learning performance of revVRNNbp (red line), bioVRNNrf (blue line), untrained RNN (gray solid line: partly out of view), and untrained RNN with connections shuffled from those learnt in bioVRNNrf (gray dotted line) when the number of RNN units was varied from 5 to 40 (horizontal axis). Leaning performance was measured by the sum of squares of differences between the state values developed at 1500-th trial by each of these four types of agent and the estimated true state values between cue and reward (vertical axis). **(F)** Mean of the elements of the recurrent and feed-forward connections (at 1500-th trial) of revVRNNbp (red line), bioVRNNrf (blue line), and untrained RNN (gray solid line).

We examined how the angle between the value weights (***w***) and the random feedback (***c***) changed across trials in bioVRNNrf. As shown in Fig. 7A, the angle was on average smaller than the chance-level angle (90°) from the beginning, while there was no further alignment over trials. This could be understood as follows. Because both the value weights (***w***) and the random feedback (***c***) were now constrained to be non-negative, these two vectors were ensured to be in a relatively close angle (i.e., in the same quadrant) from the beginning. By virtue of this loose alignment, the random feedback could act similarly to backprop-derived proper feedback, even without further alignment. We examined if the angle between the value weights (***w***) and the random feedback (***c***) at 1500-th trial was associated with the developed value of pre-reward state across simulations, but found no association ($r$ = 0.0117, $p$ = 0.908) (Fig. 7B). We then examined if the ***w-c*** angle at earlier trials (2nd - 500-th trials) was associated with the developed values at 500-th trial, with the number of simulations increased to 1000 so that small correlation could be detected. We found that the ***w-c*** angle at initial trials (2nd - around 10-th trials) was negatively correlated with the developed values of the reward state and preceding states at 500-th trial (Fig. 7C). As for the reward state, negative correlation at around 100-th - 300-th trial was also observed. These results suggest that better alignment of ***w*** and ***c*** at initial and early

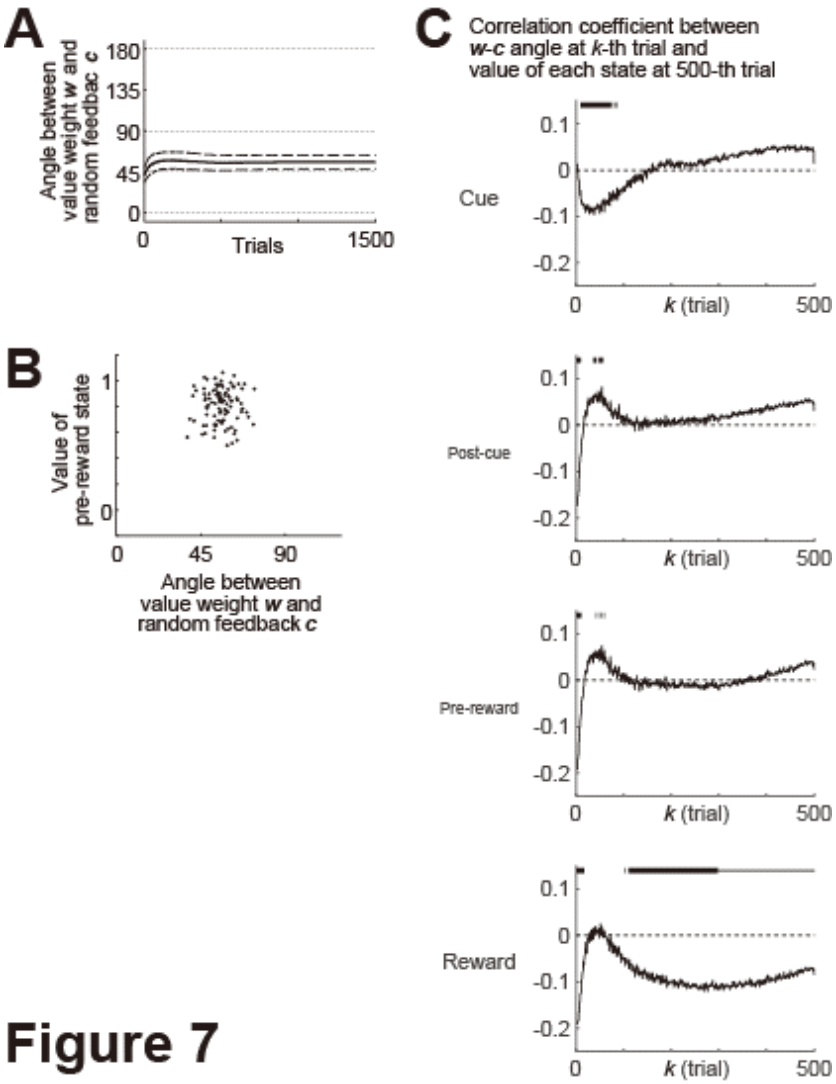


**Figure 7**

Loose alignment of the value weights (***w***) and the random feedback (***c***) in bioVRNNrf (with 12 RNN units), and its relation to the developed state values. **(A)** Over-trial changes in the angle between the value weights ***w*** and the fixed random feedback ***c***. The solid line and the dashed lines indicate the mean and ± SD across 100 simulations, respectively. **(B)** Relation between the ***w-c*** angle (horizontal axis) and the value of the pre-reward state (vertical axis) at 1500-th trial. The dots indicate the results of individual simulations. **(C)** Correlation between the ***w-c*** angle at $k$-th trial (horizontal axis) and the value of the cue, post-cue, pre-reward, or reward state (top-bottom panels) at 500-th trial across 1000 simulations. The solid lines indicate the correlation coefficient, and the short vertical bars at the top of each panel indicates the cases in which $p$-value was less than 0.05.

**Figure 7**

timings was associated with better development of state values, in line with the conjecture that loose alignment of $\boldsymbol{w}$ and $\boldsymbol{c}$ coming from the non-negative constraint supported learning. It should be noted, however, that there were cases where positive (although small) correlation was observed. Its exact reason is not sure, but it could be related to the fact that largeness of developed values or the speed of value development does not necessarily mean good learning.

We further examined how the revised value-RNN models performed in the two tasks with probabilistic structures examined above. Since the revised value-RNN models with 12 neurons appeared not able to produce the different patterns of TD-RPEs in the two tasks (TD-RPE at early reward > TD-RPE at late reward in task 1 and opposite pattern in task 2), we increased the number of neurons to 20. Then, both revVRNNbp and bioVRNNrf produced such TD-RPE patterns (Fig. 8A,B) whereas untrained RNN of both kinds (naive, and with connections shuffled from bioVRNNrf) could not (Fig. 8C,D). This indicates that the value-RNN with random feedback and further biological constraints could learn the differential characteristics of the tasks.

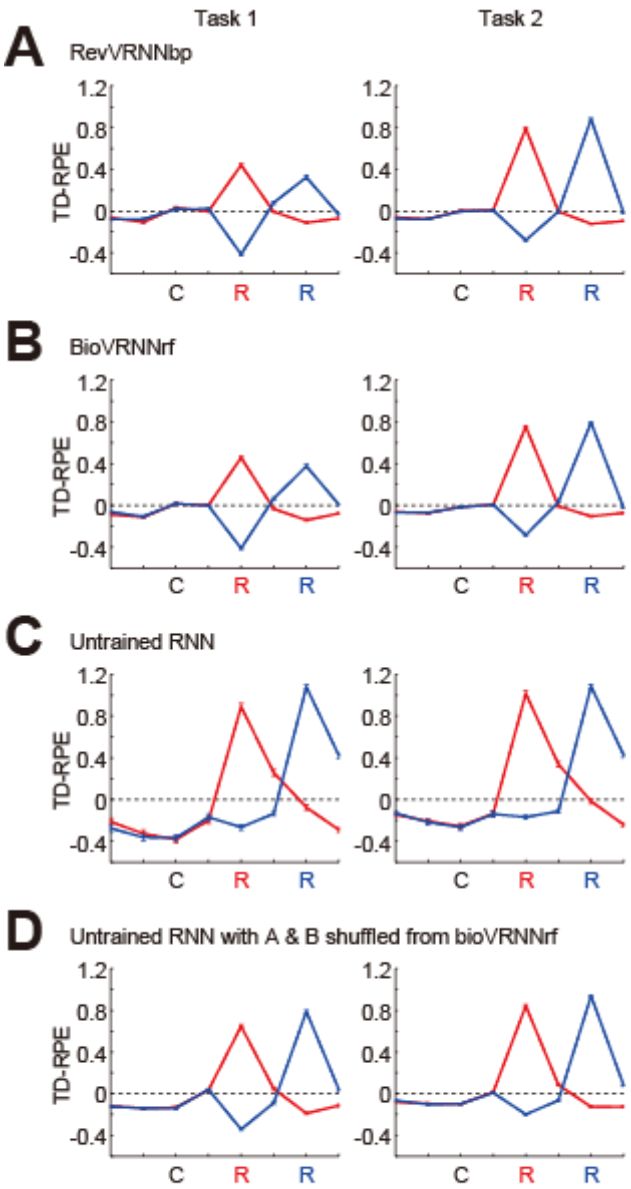


**Figure 8**

Performances of the modified value-RNN models in the two tasks having probabilistic structures, in comparison with untrained RNN having the non-negative constraint. TD-RPEs at the latest trial within 2000 trials in which reward was given at the early timing (red lines) or the late timing (blue lines) in task 1 (left) and task 2 (right), averaged across 100 simulations (error-bars indicating ± SEM across simulations), are shown for the four types of agent: (A) revVRNNbp; (B) bioVRNNrf; (C) untrained RNN with non-negative $\boldsymbol{x}$ and $\boldsymbol{w}$; (D) untrained RNN with non-negative $\boldsymbol{x}$ and $\boldsymbol{w}$ and having connections shuffled from those learnt in bioVRNNrf. The number of RNN units was 20 for all the cases.

# Discussion

We have shown that state representation and value can be learned in the RNN and its downstream by using random feedback instead of backprop-derived biologically unavailable downstream weights. In the model without non-negative constraint, the feedback alignment, previously shown for supervised learning, occurred, and we have presented an intuitive understanding of its mechanism. In the model with non-negative constraint, loose alignment occurred from the beginning because of the constraint, and it appeared to support learning. Below we discuss implementation of the value-RNN with random feedback, pointing to a crucial role of cortical DA. We also discuss relations to other proposals and suggestions, and limitations and future directions.

### *Implementation of the value-RNN with fixed random feedback, featuring a role of cortical DA*

DA neurons in the midbrain project to not only the striatum but also the cortex, especially the prefrontal cortex [37] but also the hippocampus [38]. Regarding the function of DA in the cortex, previous studies demonstrated a crucial role in working memory in the prefrontal cortex [39, 40], presumably through the effects on synaptic/ionic conductance [41, 42]. Roles of prefrontal DA in executive functions other than working memory, such as behavioral flexibility or decision making, have also been suggested [43]. Moreover, a role of hippocampal DA in modulation of aversive memory formation has been demonstrated [44]. However, although i) there has been increasing evidence that DA represents TD-RPE [45], ii) human fMRI experiments found TD-RPE correlates in cortical regions [46], iii) DAergic modulation or initiation of plasticity in the prefrontal cortex [47] or the hippocampus [48] have been demonstrated, and iv) lesion/inactivation of prefrontal or hippocampal regions were found to disrupt DA's encoding of RPE reflecting appropriate state representation [49-51], what computational role in RL is played by TD-RPE-representing DA in the cortex remains to be clarified. This is behind compared to the case of striatum, where it has been widely considered that DAergic modulation of cortico-striatal synaptic weights implements TD-RPE-based update of state/action values [6, 52].

The value-RNN with fixed random feedback and further biological constraints considered in the present work, namely the bioVRNNrf, clearly suggests a possibility that TD-RPE-representing DA modulates plasticity of RNN in the cortex so that state representation can be learnt. Different from the original value-RNN with backprop [11, 19], update of intra-cortical connections does not require downstream cortico-striatal weights but requires only non-negative fixed random feedback specific to each post-synaptic neuron. The non-negativity was assumed so that the update rule became Hebbian under positive TD-RPE, since Hebbian plasticity has been suggested for rapid plasticity of cortical synapses [36]. The fixed randomness would naturally be achieved by intrinsic heterogeneity of neurons. The successful learning performance of bioVRNNrf thus indicates that DA-dependent modulation of Hebbian plasticity of cortical (especially prefrontal and hippocampal) excitatory connections serves

for learning of state representation that captures task structure.

Notably, this suggested role of cortical DA is not mutually exclusive with the previous suggestion that prefrontal DA serves for working memory, because these two consider different effects of DA, i.e., the one on synaptic/ionic conductance and the one on synaptic plasticity. These two roles might complementarily serve for the executive functions of the prefrontal cortex.

***Relations to other proposals and suggestions***

There have been many results suggesting DA's representation of variables other than TD-RPE and/or heterogeneity of DA signals. Some of them could still be in line with the encoding of TD-RPE (e.g., [45, 53]), but others may not be easily explained by TD-RPE. Referring to such a result [54], a recent modeling work [31] proposed that DA neurons convey vector-valued error signals, which are used for supervised learning of actions in continuous space. This previous work showed that adjustment of the strengths of DA projections to the strengths of feed-forward projections, as in the backprop algorithm, was not necessary because the feedback alignment mechanism worked.

While much inspired from this previous work, the proposed implementation of the value-RNN with fixed random feedback in the present work is distinct from it in several ways, including that DA was assumed to represent TD-RPE, learning was driven by the scalar TD-RPE without vector-valued target or error signals, and probably most drastically, learning could occur by virtue of loose alignment coming from the biological constraints even without the operation of feedback alignment. Nonetheless, these two proposals can coexist. Specifically, given the heterogeneity of DA neurons, it is conceivable that there are different DA neuronal populations that respectively encode TD-RPE and vector-valued error from a target. Alternatively, same single DA neuron might represent both TD-RPE and target error depending on the context, reflecting which inputs are active in the context. This is particularly conceivable given that DA neurons receive inputs from a large number of brain regions including the cerebellum [55, 56], which has been suggested to implement supervised learning [57].

As a method to train neural networks without using backprop, the Associative Reward-Penalty ($A_{R\text{-}P}$) algorithm has been proposed [58], and applied to modeling visual-space representation in the parietal cortex [59, 60]. In the $A_{R\text{-}P}$ algorithm, the hidden units behave stochastically, and thereby the gradient could be estimated, in effect, through stochastic sampling without explicit information of the downstream weights. This mechanism is fundamentally different from the mechanisms proposed in our present work, but could simultaneously operate and achieve better performance in the actual neural systems, given that actual neurons exhibit noisy or chaotic behavior [61, 62]. This expectation could be supported by a recent work [63], which demonstrated that noise-induced learning of back projections could achieve better alignment and performance compared with the case of fixed random feedback in a feed-forward network.

### *Limitations and future directions*

We have shown that the value-RNN with fixed random feedback, and also the one with further biological constraints, could learn the value and state representation using a model with a relatively small number of simple RNN units and observation inputs in simple simulated tasks. These simplicities enabled us to derive an intuitive understanding of how the feedback alignment could occur. However, in our original models before considering the non-negativity constraint, as the number of RNN units increased, the performance of VRNNbp and VRNNrf, as well as of the untrained RNN, initially improved but then degraded when the number of RNN units increased beyond around 25. In contrast, in the original value-RNN studies [11, 19], the ability to develop belief-state-like representation was reported to improve as the number of RNN units increased to 100 or 50.

There are several possible reasons for this difference. First, as a performance measure we used the sum of squared errors between the values developed by the value-RNN and the state values estimated according to the definition, i.e., expected discounted cumulative future rewards, between cue and reward, while the previous studies focused on the similarity between the representation developed by the value-RNN and the handcrafted belief states. Second, there was a difference in the RNN unit used in the previous studies and in the present study. Specifically, as an RNN unit, we used a simple sigmoidal function. In contrast, the previous studies used the "Gated Recurrent Unit (GRU) cell" [64]. RNN with simple nonlinear unit is known to have the "vanishing gradient problem" [65]. Specifically, through repetitive learning, the gradient of the loss function becomes so small that update becomes invisible. This issue was found to be alleviated by replacing the simple RNN unit with a unit having a memory, called the Long Short-Term Memory (LSTM) unit [66]. The GRU cell was suggested to have a memory function similar to the LSTM unit [64]. We aimed at resolving the biological implausibility of the backprop method, and so we did not use the GRU or LSTM unit, whose plausibility is elusive although research has been made [67]. Third, there was also a difference in the way of weight update between the previous studies and our present study as mentioned in the Results. Specifically, while the previous studies used the BPTT [32], which considers the recursive influence of the recurrent weights in a way that lacks causality, our present study used an online learning rule, which only considers the influence of the recurrent weights at the previous time step.

Regarding biological implementation, there are still many issues, looking at the real physiology of neurons/synapses and complexity of the cortico-basal ganglia circuits. For instance, even in our biologically most constrained model, recurrent/feed-forward connections could take both positive and negative values. We think, however, that this is potentially plausible because there are both excitatory and inhibitory connections in cortical recurrent/feed-forward connections and net connection sign can be positive or negative depending on whether excitation or inhibition exceeds the other (as described in the Results and shown in Fig. 6F, the connections were on average negative in bioVRNNrf and also, less prominently, in revVRNNbp). Recent studies have shown that feed-forward and recurrent neural

networks conforming to Dale's law can perform well depending on the architecture, initialization, and update rules [68, 69]. Integration of these models and ours may be a fruitful direction. Another limitation is that our RNN unit was, while not artificially elaborated as discussed above, quite simple, operating on discrete time-steps and not having internal dynamics. Along this vein, neuronal spiking and its effects on e.g., spike-timing-dependent plasticity or synchronization were not incorporated. More specific to the cortico-basal ganglia circuit, existences of D1/D2 DA receptors and D1-direct and D2-indirect basal ganglia pathways [70-73], as well as cortical areas, layers, and cell types, were also not incorporated. Furthermore, circuit/synaptic mechanisms of how TD-RPE could be (learned to be) calculated in DA neurons (c.f., [74]) was unspecified. Future studies are expected to incorporate these factors.

# Methods

### *Value-RNN with backprop (VRNNbp)*

We constructed a value-RNN model based on the previous proposals [11, 19] but with several differences. We assumed that the activities of neurons in the RNN at time $t$+1 were determined by the activities of these neurons and neurons representing observation (cue, reward, or nothing) at time $t$:

$$\boldsymbol{x}(t+1) = \boldsymbol{f}(\mathbf{A}\boldsymbol{x}(t) + \mathbf{B}\boldsymbol{o}(t)),$$

where

$\boldsymbol{x} = (x_j)$ : activity of $j$-th neuron in the RNN ($j$ = 1, .., $n$)

$\boldsymbol{o} = (o_k)$ : activity of $k$-th neuron in the observation layer ($k$ = 1, 2)

if there was a cue at $t$, $\boldsymbol{o}(t) = (1\ 0)^{\mathrm{T}}$,

if there was a reward at $t$, $\boldsymbol{o}(t) = (0\ 1)^{\mathrm{T}}$,

and otherwise, $\boldsymbol{o}(t) = (0\ 0)^{\mathrm{T}}$

$\mathbf{A} = (A_{ij})$ : recurrent connection strength from $x_j$ to $x_i$

$\mathbf{B} = (B_{ik})$ : feed-forward connection strength from $o_k$ to $x_i$

$f(z) = 1/(1 + \exp(-z)) - 0.5$ : sigmoidal function representing neuronal input-output relation

The estimated value of the state at $t$ was calculated as

$$v(t) = \boldsymbol{w}^{\mathrm{T}}\, \boldsymbol{x}(t)$$

where

$$\boldsymbol{w} = (w_j)$$

were the value weights. The error between this estimated value and the true value, $v_{true}(t)$, was defined as:

$$\varepsilon(t) = v_{true}(t) - v(t)$$

Parameters $w_j$, $A_{ij}$, and $B_{ik}$ that minimize the squared error $\varepsilon(t)^2$ could be found by a gradient descent / error-backpropagation (backprop) method, i.e., by updating them in the directions of $-\partial(\varepsilon(t)^2)/\partial w_j$, $-\partial(\varepsilon(t)^2)/\partial A_{ij}$, and $-\partial(\varepsilon(t)^2)/\partial B_{ik}$. $-\partial(\varepsilon(t)^2)/\partial w_j$ was calculated as follows:

$$\begin{aligned}
&-\partial(\varepsilon(t)^2)/\partial w_j \\
&= -2\varepsilon(t)\partial\varepsilon(t)/\partial w_j \\
&= -2\varepsilon(t)\partial(v_{true}(t) - \boldsymbol{w}^{\mathrm{T}}\, \boldsymbol{x}(t))/\partial w_j \\
&= -2\varepsilon(t)(-x_j(t)) \\
&= 2\varepsilon(t)x_j(t) \\
&\approx 2\delta(t)x_j(t)
\end{aligned}$$

In the last line, since $\varepsilon(t)$ was unavailable as $v_{true}(t)$ was unknown, it was approximated by the TD-RPE:

$$\delta(t) = r(t) + \gamma v(t+1) - v(t).$$

$-\partial(\varepsilon(t)^2)/\partial A_{ij}$ was calculated as follows:

$$-\partial(\varepsilon(t)^2)/\partial A_{ij}$$

$= -2\varepsilon(t)\partial(v_{true}(t) - \boldsymbol{w}^{\mathrm{T}}\,\boldsymbol{x}(t))/\partial A_{ij}$

$\approx 2\delta(t)\partial(\boldsymbol{w}^{\mathrm{T}}\,\boldsymbol{f}(\mathbf{A}\boldsymbol{x}(t-1) + \mathbf{B}\boldsymbol{o}(t-1)))/\partial A_{ij}$

$= 2\delta(t)x_j(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i$

Similarly, $-\partial(\varepsilon(t)^2)/\partial B_{ik}$ was calculated as follows:

$-\partial(\varepsilon(t)^2)/\partial B_{ik}$

$\approx 2\delta(t)o_k(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i$

According to these, the update rule for the value-RNN was determined as follows:

$w_j \leftarrow w_j + a\delta(t)x_j(t)$

$A_{ij} \leftarrow A_{ij} + a\delta(t)x_j(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i$

$B_{ik} \leftarrow B_{ik} + a\delta(t)o_k(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i,$

where $a$ was the learning rate. In each simulation, the elements of **A** and **B** , as well as the elements of $\boldsymbol{x}$, were initialized to pseudo standard normal random numbers, and the elements of $\boldsymbol{w}$ were initialized to 0.

### *Value-RNN with fixed random feedback (VRNNrf)*

We considered an implementation of the value-RNN described above in the cortico-basal ganglia-DA system (Fig. 1):

$\boldsymbol{x}$ : activities of neurons in a cortical region with rich recurrent connections

**A** : recurrent connection strengths among $\boldsymbol{x}$

$\boldsymbol{o}$ : activities of neurons in a cortical region processing sensory inputs

**B** : feed-forward connection strengths from $\boldsymbol{o}$ to $\boldsymbol{x}$

$f$ : sigmoidal relationship from the input to the output of the cortical neurons

$\boldsymbol{w}$ : connection strengths from cortical neurons $\boldsymbol{x}$ to a group of striatal neurons

$v$ : activity of the group of striatal neurons

$\delta$ : activity of a group of DA neurons / released DA

The update rule for $\boldsymbol{w}$

$w_j \leftarrow w_j + a\delta(t)x_j(t)$

could be naturally implemented as cortico-striatal synaptic plasticity, which depends on DA ($\delta(t)$) and pre-synaptic (cortical) neuronal activity ($x_j(t)$). However, an issue emerged in implementation of the update rules for **A** and **B**:

$A_{ij} \leftarrow A_{ij} + a\delta(t)x_j(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i$

$B_{ik} \leftarrow B_{ik} + a\delta(t)o_k(t-1)(0.5 + x_i(t))(0.5 - x_i(t))w_i,$

Specifically, $w_i$ included in the rightmost of these update rules (for the strengths of cortico-cortical synapses $A_{ij}$ and $B_{ik}$) is the connection strength from cortical neuron $x_i$ to striatal neurons, i.e., the strength of the cortico-striatal synapses (located within the striatum), which is considered to be unavailable at the cortico-cortical synapses (located within the cortex).

As mentioned in the Introduction, this is an example of the long-standing difficulty in biological

implementation of backprop, and recently a potential solution for this difficulty, i.e., replacement of the downstream connection strengths in the update rule for upstream connections with fixed random strengths, has been demonstrated in supervised learning of feed-forward and recurrent networks [26, 30, 31]. The value-RNN, which we considered here, differed from supervised learning considered in these previous studies in two ways: i) it was TD learning, apparent in the approximation of the true error $\varepsilon(t)$ by the TD-RPE $\delta(t)$ in the derivation described above, and ii) it used a scalar error (TD-RPE) rather than a vector error. But we expected that the feedback alignment mechanism could still work at least to some extent, and explored it in this study. Specifically, we examined a modified value-RNN with fixed random feedback (VRNNrf), in which the update rules for **A** and **B** were modified as follows:

$$A_{ij} \leftarrow A_{ij} + a\delta(t)x_j(t-1)(0.5 + x_i(t))(0.5 - x_i(t))c_i$$

$$B_{ik} \leftarrow B_{ik} + a\delta(t)o_k(t-1)(0.5 + x_i(t))(0.5 - x_i(t))c_i,$$

where $w_i$ in the update rules of the value-RNN with backprop (VRNNbp) was replaced with a fixed random parameter $c_i$. Notably, these modified update rules for the cortico-cortical connections **A** and **B** required only pre-synaptic activities ($x_j(t-1)$, $o_k(t-1)$), post-synaptic activities ($x_i(t)$), TD-RPE-representing DA ($\delta(t)$), and fixed random strengths ($c_i$), which would all be available at the cortico-cortical synapses given that VTA DA neurons project not only to the striatum but also to the cortex and random $c_i$ could be provided by intrinsic heterogeneity. In each simulation, the elements of $\boldsymbol{c}$ were initialized to pseudo standard normal random numbers.

### *Revised value-RNN models with further biological constraints*

In the later part of this study, we examined revised value-RNN models with further biological constraints. Specifically, we considered models, in which the value weights and the activities of neurons in the RNN were constrained to be non-negative. In order to do so, the update rule for $\boldsymbol{w}$ was modified to:

$$w_j \leftarrow \max(0, w_j + a\delta(t)x_j(t)),$$

where $\max(q_1, q_2)$ returned the maximum of $q_1$ and $q_2$. Also, the sigmoidal input-output function was replaced with

$$f(z) = 1/(1 + \exp(-z)),$$

and the elements of $\boldsymbol{x}$ were initialized to pseudo uniform [0 1] random numbers. The backprop-based update rules for **A** and **B** in VRNNbp were replaced with

$$A_{ij} \leftarrow A_{ij} + a\delta(t)x_j(t-1)x_i(t)(1 - x_i(t))w_i$$

$$B_{ik} \leftarrow B_{ik} + a\delta(t)o_k(t-1)x_i(t)(1 - x_i(t))w_i.$$

We referred to the model with these modifications to VRNNbp as revVRNNbp.

As a revised value-RNN with fixed random feedback (VRNNrf), in addition to the abovementioned modifications of the update of $\boldsymbol{w}$, the sigmoidal input-output function, and the initialization of $\boldsymbol{x}$, the fixed random feedback $\boldsymbol{c}$ was assumed to be non-negative. Specifically, the elements of $\boldsymbol{c}$ were set to pseudo uniform [0 1] random numbers. Moreover, the update rules for **A** and

**B** were replaced with

(when $x_i(t) \leq 0.5$)

$$A_{ij} \leftarrow A_{ij} + a\delta(t)x_j(t-1)x_i(t)(1 - x_i(t))c_i$$

$$B_{ik} \leftarrow B_{ik} + a\delta(t)o_k(t-1)x_i(t)(1 - x_i(t))c_i.$$

(when $x_i(t) > 0.5$)

$$A_{ij} \leftarrow A_{ij} + 0.25a\delta(t)x_j(t-1)c_i$$

$$B_{ik} \leftarrow B_{ik} + 0.25a\delta(t)o_k(t-1)c_i.$$

so that the originally non-monotonic dependence on $x_i(t)$ (post-synaptic activity) became monotonic + saturation (Fig. 5B). These update rules with non-negative $c_i$ could be said to be Hebbian with additional modulation by TD-RPE (Hebbian under positive TD-RPE). We referred to the model with these modifications to VRNNrf as bioVRNNrf.

### *Simulation of the tasks*

In the Pavlovian cue-reward association task, at time 1 of each trial, cue observation was received by the RNN, and at time 4, reward observation was received. Trial was pseudo-randomly ended at time 7, 8, 9, or 10, and the next trial started from the next time-step. Reward size was $r = 1$. The tasks with probabilistic structures (task 1 and task 2) were implemented in the same way except that reward timing was not time 4 but time 3 or 5 with equal probabilities, specifically, 50% and 50% in task 1 and 30% and 30% in task 2, and there was no reward in the remaining 40% of trials in task 2.

The cue or reward state/timing, mentioned in the text and marked in the figures, was defined to be the timing when the RNN received the cue or reward observation, respectively. Specifically, if $\boldsymbol{o}(t) = (1\ 0)^T$ or $\boldsymbol{o}(t) = (0\ 1)^T$ at time $t$, $t + 1$ was defined to be a cue or reward timing, respectively. For the agents with punctate (CSC) representation, each timing in the tasks was represented by a 10-dimensional one-hot vector, starting from $(1\ 0\ 0\ ...\ 0)^T$ for the cue state, with the next state $(0\ 1\ 0\ ...\ 0)^T$ and so on.

Unless otherwise mentioned, parameters were set to the following values. Learning rate ($a$): 0.1 (normalization by the squared norm of feature vector was not implemented). Time discount factor ($\gamma$): 0.8.

### *Estimation of true state values*

As for the Pavlovian cue-reward association task, we defined states by relative timings from the cue, and estimated their (true) state values by simulations according to the definition of state value. Specifically, we generated a sequence of cues and rewards corresponding to 1000 trials, and calculated cumulative discounted future rewards within the sequence:

$$\Sigma_{t_\mathrm{rew}}(r\gamma^{t-\mathrm{rew}}),$$

where $t_{rew}$ denotes the time-step of each reward counted from the starting state, starting from -2, -1, ..., and +6 time steps from a cue. We repeated this 1000 times, generating 1000 sequences (i.e., 1000 simulations of 1000 trials), with different sets of pseudo-random numbers, and calculated an average over these 1000 sequences so as to estimate the expected cumulative discounted future rewards, i.e., state value (by definition) for each state (-2, -1, ..., and +6 time steps from cue). Using these estimated true state values, we calculated TD-RPE at each state (-2, -1, ..., and +5 time steps from cue).

In a similar manner, we defined states and estimated true state values, and also calculated TD-RPE, for tasks 1 and 2 that had probabilistic structures. As for task 1, we defined the following states: -2, -1, .., and +2 time steps from cue (i.e., states visited (entered) before knowing whether reward was given at the early timing (= +2 time step from cue)), +3, 4, 5, and 6 time steps from cue after reception of reward at the early timing, and +3, 4, 5, and 6 time steps from cue after no reception of reward at the early timing (in total 5 + 4 + 4 = 13 states) (Fig. 4Bc, left-top). We generated 10000 sequences of cues and rewards corresponding to 1000 trials (i.e., 10000 simulations of 1000 trials), and for each state, calculated cumulative discounted future rewards within the sequence for each of 10000 simulations and took an average to obtain the expected cumulative discounted future rewards (i.e., estimation of state value) (Fig. 4Bc, left-bottom). Using the estimated state values, we calculated TD-RPE (Fig. 4C, left).

As for task 2, we defined the following states: -2, -1, .., and +2 time steps from cue (i.e., states visited (entered) before knowing whether reward was given at the early timing), +3, 4, 5, and 6 time steps from cue after reception of reward at the early timing, +3 and 4 time steps from cue after no reception of reward at the early timing (states visited (entered) before knowing whether reward was given at the late timing (= +4 time step from cue)), +5 and 6 time steps from cue after reception of reward at the late timing, and +5 and 6 time steps from cue after no reception of reward at both early and late timings (in total 5 + 4 + 2 + 2 + 2 = 15 states) (Fig. 4Bc, right-top). We estimated state values of these states (Fig. 4Bc, right-bottom), and also calculated TD-RPE (Fig. 4C, right), in similar manners to the above.

### *Analyses, software, and code availability*

SEM (Standard error of the mean) was approximated by SD (standard deviation)/$\sqrt{N}$ (number of samples). Linear regression and principal component analysis (PCA) were conducted by using R (functions lm and prcomp). Simulations were conducted by using MATLAB, and pseudo-random numbers were implemented by using rand, randn, and randperm functions. All the codes will be made available at GitHub upon publication of this work in a journal.

## References


1. Montague, P.R., Dayan, P. & Sejnowski, T.J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* **16**, 1936-1947 (1996).
2. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).
3. Niv, Y. & Schoenbaum, G. Dialogues on prediction errors. *Trends Cogn Sci* **12**, 265-272 (2008).
4. Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85-88 (2012).
5. Steinberg, E.E.*, et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* **16**, 966-973 (2013).
6. Reynolds, J.N., Hyland, B.I. & Wickens, J.R. A cellular mechanism of reward-related learning. *Nature* **413**, 67-70 (2001).
7. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D.J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848-851 (2008).
8. Yagishita, S.*, et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616-1620 (2014).
9. Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J. & Daw, N.D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput Biol* **13**, e1005768 (2017).
10. Stachenfeld, K.L., Botvinick, M.M. & Gershman, S.J. The hippocampus as a predictive map. *Nat Neurosci* **20**, 1643-1653 (2017).
11. Qian, L.*, et al.* The role of prospective contingency in the control of behavior and dopamine signals during associative learning. *bioRxiv* (2024).
12. Langdon, A.J., Sharpe, M.J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr Opin Neurobiol* **49**, 1-7 (2018).
13. Redish, A.D., Jensen, S., Johnson, A. & Kurth-Nelson, Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol Rev* **114**, 784-805 (2007).
14. Gershman, S.J., Jones, C.E., Norman, K.A., Monfils, M.H. & Niv, Y. Gradual extinction prevents the return of fear: implications for the discovery of state. *Front Behav Neurosci* **7**, 164 (2013).
15. Shimomura, K., Kato, A. & Morita, K. Rigid reduced successor representation as a potential mechanism for addiction. *Eur J Neurosci* **53**, 3768-3790 (2021).
16. Feng, Z., Nagase, A.M. & Morita, K. A Reinforcement Learning Approach to Understanding Procrastination: Does Inaccurate Value Approximation Cause Irrational Postponing of a Task? *Front Neurosci* **15**, 660595 (2021).
17. Sato, R., Shimomura, K. & Morita, K. Opponent learning with different representations in the cortico-basal ganglia pathways can develop obsession-compulsion cycle. *PLoS Comput Biol* **19**, e1011206 (2023).
18. Niv, Y. Learning task-state representations. *Nat Neurosci* **22**, 1544-1553 (2019).
19. Hennig, J.A.*, et al.* Emergence of belief-like representations through reinforcement learning.

*PLoS Comput Biol* **19**, e1011067 (2023).
20. Amari, S. A Theory of Adaptive Pattern Classifiers. *IEEE Transactions on Electronic Computers* **EC-16**, 299-307 (1967).
21. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. Learning representations by back-propagating errors. *Nature* **323**, 533-536 (1986).
22. Doya, K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr Opin Neurobiol* **10**, 732-739 (2000).
23. O'Doherty, J.*, et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452-454 (2004).
24. Grossberg, S. Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science* **11**, 23-63 (1987).
25. Crick, F. The recent excitement about neural networks. *Nature* **337**, 129-132 (1989).
26. Lillicrap, T.P., Cownden, D., Tweed, D.B. & Akerman, C.J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat Commun* **7**, 13276 (2016).
27. Whittington, J.C.R. & Bogacz, R. Theories of Error Back-Propagation in the Brain. *Trends Cogn Sci* **23**, 235-250 (2019).
28. Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J. & Hinton, G. Backpropagation and the brain. *Nat Rev Neurosci* **21**, 335-346 (2020).
29. Song, Y.*, et al.* Inferring neural activity before plasticity as a foundation for learning beyond backpropagation. *Nat Neurosci* **27**, 348-358 (2024).
30. Murray, J.M. Local online learning in recurrent networks with random feedback. *Elife* **8** (2019).
31. Wärnberg, E. & Kumar, A. Feasibility of dopamine as a vector-valued feedback signal in the basal ganglia. *Proc Natl Acad Sci U S A* **120**, e2221994120 (2023).
32. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. Learning Internal Representations by Error Propagation. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1 Foundations* (ed. D.E. Rumelhart, McClelland, J.L., The PDP Group) 318-362 (MIT Press, Cambridge, 1985).
33. Ludvig, E.A., Sutton, R.S. & Kehoe, E.J. Evaluating the TD model of classical conditioning. *Learn Behav* **40**, 305-319 (2012).
34. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction (Second Edition)* (MIT Press, Cambridge, MA, 2018).
35. Starkweather, C.K., Babayan, B.M., Uchida, N. & Gershman, S.J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**, 581-589 (2017).
36. Feldman, D.E. Synaptic mechanisms for plasticity in neocortex. *Annu Rev Neurosci* **32**, 33-55 (2009).
37. Williams, S.M. & Goldman-Rakic, P.S. Widespread origin of the primate mesofrontal dopamine system. *Cereb Cortex* **8**, 321-345 (1998).
38. Broussard, J.I.*, et al.* Dopamine Regulates Aversive Contextual Learning and Associated In Vivo Synaptic Plasticity in the Hippocampus. *Cell Rep* **14**, 1930-1939 (2016).
39. Brozoski, T.J., Brown, R.M., Rosvold, H.E. & Goldman, P.S. Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey. *Science* **205**, 929-932 (1979).
40. Sawaguchi, T. & Goldman-Rakic, P.S. D1 dopamine receptors in prefrontal cortex: involvement

in working memory. *Science* **251**, 947-950 (1991).
41. Durstewitz, D., Seamans, J.K. & Sejnowski, T.J. Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *J Neurophysiol* **83**, 1733-1750 (2000).
42. Brunel, N. & Wang, X. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J Comput Neurosci* **11**, 63-85 (2001).
43. Floresco, S.B. & Magyar, O. Mesocortical dopamine modulation of executive functions: beyond working memory. *Psychopharmacology (Berl)* **188**, 567-585 (2006).
44. Tsetsenis, T.*, et al.* Midbrain dopaminergic innervation of the hippocampus is sufficient to modulate formation of aversive memories. *Proc Natl Acad Sci U S A* **118**, e2111069118 (2021).
45. Kim, H.R.*, et al.* A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**, 1600-1616.e1625 (2020).
46. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329-337 (2003).
47. Otani, S., Daniel, H., Roisin, M.P. & Crepel, F. Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cereb Cortex* **13**, 1251-1256 (2003).
48. Sayegh, F.J.P.*, et al.* Ventral tegmental area dopamine projections to the hippocampus trigger long-term potentiation and contextual learning. *Nat Commun* **15**, 4100 (2024).
49. Takahashi, Y.K.*, et al.* Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat Neurosci* **14**, 1590-1597 (2011).
50. Starkweather, C.K., Gershman, S.J. & Uchida, N. The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98**, 616-629.e616 (2018).
51. Takahashi, Y.K.*, et al.* Expectancy-related changes in firing of dopamine neurons depend on hippocampus. *bioRxiv* **https://doi.org/10.1101/2023.07.19.549728** (2023).
52. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337-1340 (2005).
53. Lee, R.S., Engelhard, B., Witten, I.B. & Daw, N.D. A vector reward prediction error model explains dopaminergic heterogeneity. *bioRxiv* **https://doi.org/10.1101/2022.02.28.482379** (2022).
54. Avvisati, R.*, et al.* Distributional coding of associative learning in discrete populations of midbrain dopamine neurons. *Cell Rep* **43**, 114080 (2024).
55. Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A. & Uchida, N. Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron* **74**, 858-873 (2012).
56. Carta, I., Chen, C.H., Schott, A.L., Dorizan, S. & Khodakhah, K. Cerebellar modulation of the reward circuitry and social behavior. *Science* **363** (2019).
57. Marr, D. A theory of cerebellar cortex. *J Physiol* **202**, 437-470 (1969).
58. Barto, A.G. & Jordan, M.I. Gradient Following Without Back-Propagation in Layered Networks. in *Proceedings of the First Annual International Conference on Neural Networks Vol. II* 629-636 (San Diego, CA., 1987).
59. Mazzoni, P., Andersen, R.A. & Jordan, M.I. A more biologically plausible learning rule for neural networks. *Proc Natl Acad Sci U S A* **88**, 4433-4437 (1991).
60. Mazzoni, P., Andersen, R.A. & Jordan, M.I. A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a. *Cereb Cortex* **1**, 293-307 (1991).
61. Faisal, A.A., Selen, L.P. & Wolpert, D.M. Noise in the nervous system. *Nat Rev Neurosci* **9**, 292-

303 (2008).
62. Aihara, K. & Matsumoto, G. Chaotic oscillations and bifurcations in squid giant axons. in *Chaos* (ed. A.V. Holden) (Princeton University Press, 1986).
63. Max, K.*, et al.* Learning efficient backprojections across cortical hierarchies in real time. *Nature Machine Intelligence* **6**, 619–630 (2024).
64. Cho, K.*, et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *arXiv* **1406.1078** (2014).
65. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **6**, 107-115 (1998).
66. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735-1780 (1997).
67. Costa, R.P., Assael, Y.M., Shillingford, B., De Freitas, N. & Vogels, T. Cortical microcircuits as gated-recurrent neural networks. *Advances in Neural Information Processing Systems* (2017).
68. Cornford, J.*, et al.* Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *bioRxiv* **https://doi.org/10.1101/2020.11.02.364968** (2021).
69. Li, P., Cornford, J., Ghosh, A. & Richards, B. Learning better with Dale's Law: A Spectral Perspective. *bioRxiv* **https://doi.org/10.1101/2023.06.28.546924** (2023).
70. Gerfen, C.R. & Surmeier, D.J. Modulation of Striatal Projection Systems by Dopamine. *Annu Rev Neurosci* **34**, 441-466 (2011).
71. Collins, A.G. & Frank, M.J. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol Rev* **121**, 337-366 (2014).
72. Mikhael, J.G. & Bogacz, R. Learning Reward Uncertainty in the Basal Ganglia. *PLoS Comput Biol* **12**, e1005062 (2016).
73. Lowet, A.S.*, et al.* An opponent striatal circuit for distributional reinforcement learning. *bioRxiv* **https://doi.org/10.1101/2024.01.02.573966** (2024).
74. Cone, I., Clopath, C. & Shouval, H.Z. Learning to express reward prediction error-like dopaminergic activity requires plastic representations of time. *Nat Commun* **15**, 5856 (2024).