frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fncom.2020.00066

Computational Characteristics of the Striatal Dopamine System Described by Reinforcement Learning With Fast Generalization

Yoshihisa Fujita[1]*, *Sho Yagishita*[2,3], *Haruo Kasai*[2,3] *and Shin Ishii*[1,3,4]

[1] Integrated Systems Biology Laboratory, Department of Systems Science, Graduate School of Informatics, Kyoto University, Kyoto, Japan, [2] Laboratory of Structural Physiology, Center for Disease Biology and Integrative Medicine, Faculty of Medicine, The University of Tokyo, Tokyo, Japan, [3] International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Tokyo, Japan, [4] Neural Information Processing Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan

Generalization is the ability to apply past experience to similar but non-identical situations. It not only affects stimulus-outcome relationships, as observed in conditioning experiments, but may also be essential for adaptive behaviors, which involve the interaction between individuals and their environment. Computational modeling could potentially clarify the effect of generalization on adaptive behaviors and how this effect emerges from the underlying computation. Recent neurobiological observation indicated that the striatal dopamine system achieves generalization and subsequent discrimination by updating the corticostriatal synaptic connections in differential response to reward and punishment. In this study, we analyzed how computational characteristics in this neurobiological system affects adaptive behaviors. We proposed a novel reinforcement learning model with multilayer neural networks in which the synaptic weights of only the last layer are updated according to the prediction error. We set fixed connections between the input and hidden layers to maintain the similarity of inputs in the hidden-layer representation. This network enabled fast generalization of reward and punishment learning, and thereby facilitated safe and efficient exploration of spatial navigation tasks. Notably, it demonstrated a quick reward approach and efficient punishment aversion in the early learning phase, compared to algorithms that do not show generalization. However, disturbance of the network that causes noisy generalization and impaired discrimination induced maladaptive valuation. These results suggested the advantage and potential drawback of computation by the striatal dopamine system with regard to adaptive behaviors.

Keywords: generalization, adaptive behaviors, reward learning, striatum, dopamine-dependent plasticity, reinforcement learning, artificial neural networks

OPEN ACCESS

Edited by:
Uma Shanker Tiwary,
Indian Institute of Information Technology, Allahabad, India

Reviewed by:
Adam Ponzi,
Okinawa Institute of Science and Technology Graduate University, Japan
Alekhya Mandali,
University of Cambridge, United Kingdom

****Correspondence:***
Yoshihisa Fujita
fujita.yoshihisa.4s@kyoto-u.ac.jp

Received: *11 March 2020*
Accepted: *08 June 2020*
Published: *22 July 2020*

Citation:
Fujita Y, Yagishita S, Kasai H and Ishii S (2020) Computational Characteristics of the Striatal Dopamine System Described by Reinforcement Learning With Fast Generalization. Front. Comput. Neurosci. 14:66. doi: 10.3389/fncom.2020.00066

INTRODUCTION

Animals' survival incorporates reward-seeking behavior accompanied by risks. Outcome observation resulting from the pairing of a current state and a taken action provide clues to ensure optimal behaviors, but it may be associated with substantial energy consumption and aversive experiences. Such a learning process is inefficient and even harmful, especially when animals are required to adapt to new environments. Animals instead generalize their previous experiences to predict outcome, even in novel situations. If the prediction due to generalization is different from the actual observation, the prediction is then reshaped by discrimination learning. Generalization and discrimination may be essential for efficient adaptive behaviors, whereas abnormalities in these functions can be maladaptive. Generalization to an abnormal extent has been implicated in psychiatric disorders (Dunsmoor and Paz, 2015; Kahnt and Tobler, 2016; Asok et al., 2018). A recent neurobiological study presented the possibility that disrupted discrimination is involved in psychotic symptoms (Iino et al., 2020).

Psychological studies have investigated generalization by using the conditioning paradigm in behavioral experiments (Ghirlanda and Enquist, 2003). If a response has been established by a stimulus paired with an outcome (i.e., reward or punishment), then resembling stimuli evoke similar responses. This "law of effect" depends on the extent that the second stimulus resembles the first stimulus, and is termed "stimulus generalization" (Thorndike, 1898; Ghirlanda and Enquist, 2003). Discrimination can then occur if the first stimulus is paired with a reward but the resembling stimulus is paired with no reward; as a consequence, only the first stimulus elicits a response. How the brain establishes stimulus generalization has been explained, based on artificial neural networks (Shepard and Kannappan, 1991; Ghirlanda and Enquist, 1998; Franks and Ruxton, 2008; Wisniewski et al., 2012). These previous studies have focused on the stimulus–response relationship, although generalization can be incorporated in the interaction between individuals and the environment. Reinforcement learning involves this interaction and is used as a model of reward-driven learning (Frank et al., 2004; Doya, 2007; Glimcher, 2011). In the field of artificial intelligence, reinforcement learning in combination with artificial neural networks achieves a high performance, which suggests the contribution of generalization to adaptive behaviors (Mnih et al., 2015). However, neurobiological evidence indicates that the neural system has a unique computation for reward-driven learning and generalization, compared to algorithms used for artificial intelligence (Whittington and Bogacz, 2019). This raises the question regarding how its computational characteristics differ from those of ordinary algorithms. Advantages and shortcomings should exist with regard to adaptive behaviors.

Theoretical attempts have focused on separate systems that process positive and negative values in the brain (DeLong, 1990; Nambu, 2007; Amemori et al., 2011; Collins and Frank, 2014), which are not adopted when using ordinary reinforcement learning. Dopamine and its main target area, the striatum, play central roles in reward- and punishment-related learning (Meredith et al., 2008). Dopaminergic neurons show a positive response to a greater-than-expected reward and a negative response to a less-than-expected reward, which indicates that dopamine codes reward prediction error (Schultz, 2015). The striatum receives dopamine signals and glutamatergic input from the cortex and thalamus (Tepper et al., 2007). Dopamine modulates synaptic plasticity between the cortex and the striatum during reward-related learning (Reynolds et al., 2001). The striatum is primarily composed of spiny projection neurons (SPNs), which can be divided into SPNs that primarily express the dopamine D1 receptor (D1-SPNs) and SPNs that express the dopamine D2 receptor (D2-SPNs) (Surmeier et al., 2010). D1-SPNs respond to phasic increases in dopamine (Yagishita et al., 2014), whereas D2-SPNs respond to the phasic decreases in dopamine (Hikida et al., 2013; Iino et al., 2020). Perturbing the activity of D1- and D2-SPNs inhibits reward learning and punishment learning, respectively (Hikida et al., 2013). The advantages of having such separate systems have been discussed in computational studies (Mikhael and Bogacz, 2016; Elfwing and Seymour, 2017). One study (Mikhael and Bogacz, 2016) demonstrated an advantage in learning reward uncertainty. Another study (Elfwing and Seymour, 2017) showed the possibility of achieving safe behaviors.

However, recent neural recordings and optogenetic manipulations provide some data that suggest different roles of SPNs from those in the existing models (Cox and Witten, 2019). Our recent experiments with a classical conditioning task found that D1- and D2-SPNs are differentially responsible for stimulus generalization and discrimination (Iino et al., 2020). The same series of experiments revealed that stimulus generalization/discrimination occurred solely by dopamine-dependent plasticity of SPN spines that receive cortical inputs, which implies that updating in other connections (e.g., intracortical synaptic connections) would be minor. These observations suggest learning rules that update the synaptic weights of the last layer in a multilayer neural network—in our case, corticostriatal connections—are essential. Such learning rules actually enable remarkably fast learning in reservoir computing (RC) and extreme learning machine (ELM) (Maass et al., 2002; Huang et al., 2006; Lukoševičius and Jaeger, 2009). RC and ELM are neural networks that train only their readouts (i.e., the last layer connections); RC has recurrent connections, whereas ELM does not. Iino et al. also showed that administration of methamphetamine, which causes psychosis (e.g., delusions) in humans, impaired discrimination function in mice by altering dopamine dynamics associated with unexpected reward omission (Iino et al., 2020). Taken together with the physiological functions of D2-SPN described above, these findings suggest that impairment of dopamine-dependent corticostriatal plasticity of D2-SPNs can induce abnormal value prediction via disrupted discrimination.

In this study, we propose a novel reinforcement learning model that reproduces stimulus generalization and discrimination, while accounting for the physiological characteristics of striatal SPNs. We used a neural network for value estimation and introduced fixed connections between the input and hidden layers, as in RC and ELM. RC has the potential for context-dependent value estimation because of its recurrent connections. However, for simplicity, we adopted a feed-forward neural network with the same architecture as ELM. The extent of stimulus generalization depends on the similarity of stimuli; therefore, we did not use random and fixed connections as implemented in ordinary ELM. We instead set the fixed connections so as to maintain the similarity of inputs in the hidden-layer representation. In addition, we introduced connections that had weights separately updated by positive and negative reward prediction errors, while taking into account the differential roles of D1- and D2-SPNs in the striatum. This neural network enabled generalization in a quick manner. We named this model "Outspread Valuation for Reward Learning and Punishment Learning" ("OVaRLAP"). Using this model for painful grid-world navigation tasks, we first evaluated the contribution of generalization and discrimination to adaptive behaviors. The OVaRLAP model performed safe and efficient exploration, suggesting that quick generalization of punishment learning contributed to safe and efficient reward-seeking and pain-avoiding. We then tested disturbed OVaRLAP in a painless grid-world to examine whether abnormal generalization and discrimination underlies maladaptive behaviors, as implied in psychological and psychiatric studies (Buss and Daniell, 1967; Ralph, 1968; Kahnt and Tobler, 2016). We introduced impairment of learning from negative prediction error, which disables discrimination, based on the physiological findings (Iino et al., 2020). We found that impaired discrimination combined with noisy generalization induced aberrant valuation, after repeating reward-seeking behaviors. These results showed that the unique computation suggested by the striatal dopamine system facilitated safe and efficient exploration, but on the other hand had potential defects which can cause maladaptive behaviors.

METHODS

We developed OVaRLAP to analyze how the computational characteristics in the striatal dopamine system affect behaviors. We first evaluated OVaRLAP in a spatial navigation task in painful grid-worlds by comparing its performance with those of two other representative algorithms (**Figures 1A,B**). We then examined the behavior of disturbed OVaRLAP in a spatial navigation task in a painless grid-world (**Figures 1C,D**).

Model Description of OVaRLAP

A neural network was used for value estimation in OVaRLAP. It consists of an input layer, a hidden layer, two pre-output neurons, and an output neuron (**Figure 1A**, left). The input represents the state at discretized time, t. We applied our learning method to two types of navigation tasks, both in two-dimensional grid-worlds, in which position (x, y), $x \in \{1, \ldots, 20\}$, and $y \in \{1, \ldots 20\}$ were simply represented by a two-dimensional index function:

$$I_{ij}(x, y, t) = \begin{cases} 1, \text{if } i = x \text{ and } j = y \\ 0, \quad \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, 20$ and $j = 1, \ldots, 20$.

The input signals were transformed by fixed connections into hidden-layer activity patterns. We set the number of hidden-layer neurons and the weight of fixed connections so as to reproduce the shape of the generalization gradient, which has been commonly observed across various species, behavioral contexts, and sensory modalities (Ghirlanda and Enquist, 2003). The number of hidden-layer neurons was set to 900. The hidden-layer activity $h_k(x, y, t)$ for $k = 1, \ldots, 900$ was given by

$$h_k(x, y, t) = \sum_i \sum_j M_{ij,k} I_{ij}(x, y, t)$$

in which $M_{ij,k}$ is the weight of the fixed connection from the input-layer neuron (i, j) to the hidden-layer neuron k. Based on the definition of $I_{ij}(x, y, t)$, $h_k(x, y, t)$ was also represented as

$$h_k\left(x, y, t\right) = M_{xy,k}$$

in which (x, y) is the agent's position at time t.

For the generalization capability, we assumed the hidden-layer representation becomes similar when the input signal is similar. In our case, the values of $h_k(x_1, y_1, t)$ and $h_k(x_2, y_2, t)$ are similar if position (x_1, y_1) is close to position (x_2, y_2), based on Euclidian distance. To reflect this request, we set the fixed connections $M_{ij,k}$ to follow the two-dimensional Gaussian function:

$$M_{ij,k} = \frac{\exp\left\{-\left(\frac{(i-a_k)^2}{2\sigma_k^2} + \frac{(j-b_k)^2}{2\sigma_k^2}\right)\right\} + \varepsilon_{ij,k}}{D}$$

in which D is the number of the input-layer neurons used for normalization ($D = 400$, because of the 20 × 20 grid-worlds); a_k and b_k denote the center and are set as $a_k \in \{1, \ldots, 20\}$, $b_k \in \{1, \ldots, 20\}$, and $20\,(a_k - 1) + b_k = ceil(400k/900)$ so that the central location (a_k, b_k) in the 20×20 2D space was linearly correlated with the location $(\tilde{a}_k, \tilde{b}_k)$ in 30×30 2D space where $\tilde{a}_k = 1 + floor((k-1)/30)$ and $\tilde{b}_k = mod(k-1,\ 30) + 1$; σ_k^2 is the variance sampled from a log-normal distribution with the metaparameter θ as follows:

$$\sigma_k^2 \sim \text{LogN}\left(-0.7/\theta,\ 0.7\theta\right);$$

and $\varepsilon_{ij,k}$ is the noise, as defined below. The distribution of σ_k^2 resulted that the hidden-layer neurons ranged from neurons that responded to specific input to neurons that responded to a wide range of input. This setting is consistent with physiological observations (Bordi and LeDoux, 1992) and underlies generalization in OVaRLAP. We set the metaparameter θ to regulate the amount of generalization. The noise $\varepsilon_{ij,k}$ is unnecessary in normal cases; however, we introduced it for the purpose of analyzing the effect of abnormal perturbation in OVaRLAP (**Figure 1C**). When simulating abnormal perturbation, we applied noise $\varepsilon_{ij,k}$ by using the following formulas:

$$\text{P}\left(\varepsilon_{ij,k} = A\right) = \rho$$

and

$$\text{P}\left(\varepsilon_{ij,k} = 0\right) = 1 - \rho.$$

A denotes the noise strength. Apparently, no noise exists if A or ρ is zero. This noise was initially introduced but not changed

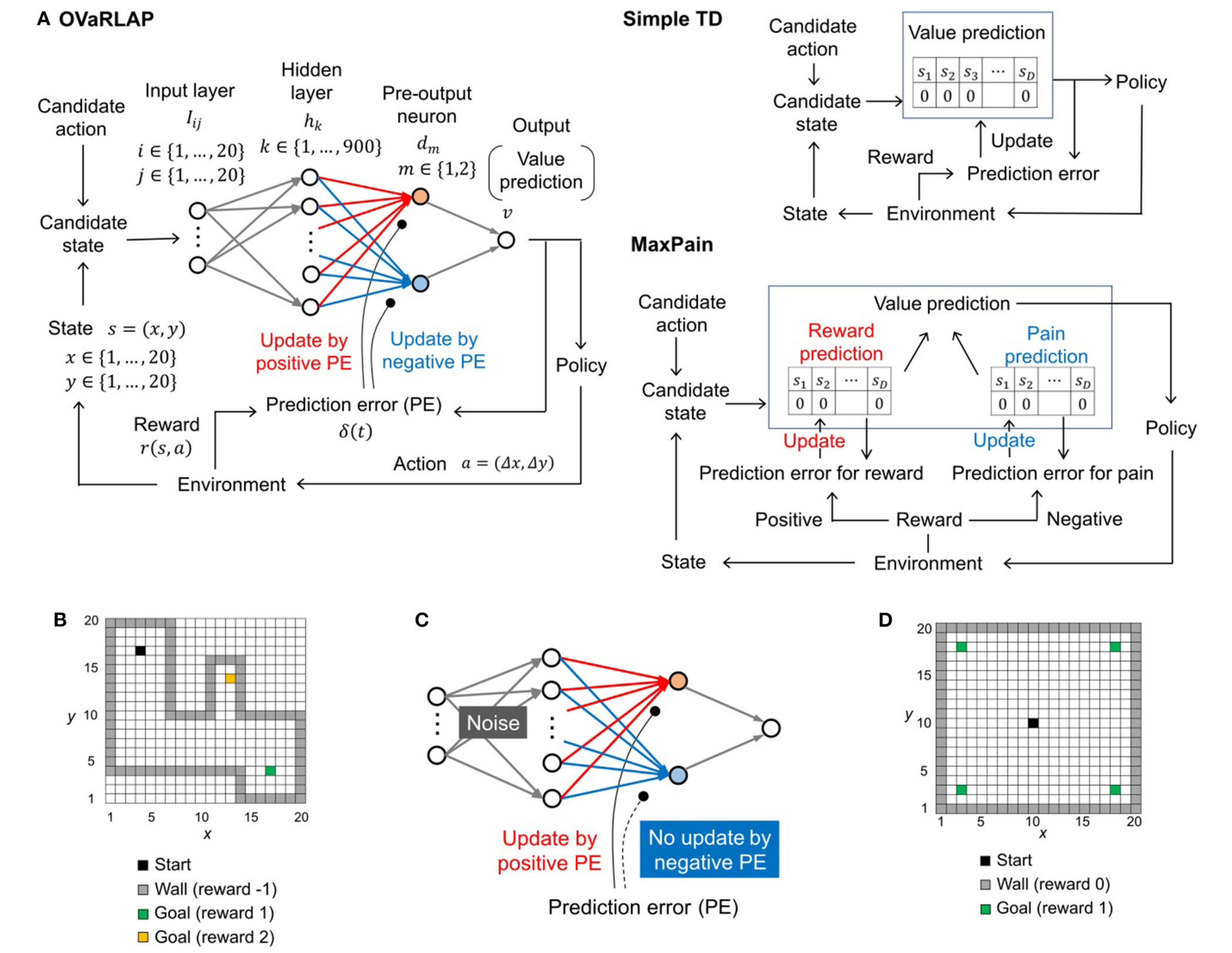

FIGURE 1 | The architecture of TD learning algorithms and the task. **(A)** Schematic diagrams of OVaRLAP (left), simple TD learning (middle), and MaxPain (right). In OVaRLAP, the value prediction used a neural network with fixed connections (gray arrows) and distinct connections updated by positive and negative TD errors (red and blue arrows, respectively). In our implementations of simple TD learning and MaxPain, the predicted value is represented by look-up tables in which the state, s_k, indicates a position (x, y) in the two-dimensional grid-world. **(B)** The navigation task in the two-dimensional grid-world. The non-gray squares and gray squares indicate passable and not passable, respectively. The black square is the starting position, and the green and yellow squares are the goals at which the agent receives the positive rewards of 1 and 2, respectively. When the agent receives a positive reward, an episode ends so that the agent restarts the task from the starting position. If an agent hits a wall (i.e., a not passable square), it receives a negative reward of -1, but continues the task by staying at the same square. **(C)** Schematic diagram of OVaRLAP in which updating connections, based on the negative TD error (blue connections), are impaired. In addition, noise is introduced to induce an anomaly in the initialization of the fixed connections between the input and the hidden layers (for the details of noise here, see Section Model Description of OVaRLAP). **(D)** The navigation task in the two-dimensional grid-world. The black square is the starting position. The green squares are the goals at which the agent receives a positive reward of one. If the agent receives a positive reward, a single episode ends, and a new episode restarts from the starting position. In this task, when the agent hits a wall (i.e., a gray square), no negative reward is given, and the agent remains at the same square.

through the learning process. In our simulation experiment, we applied different realizations of noise $\varepsilon_{ij,k}$ under specific values of A and ρ, and examined the collective behaviors of the learning.

The two pre-output neurons then received signals from the hidden-layer neurons. For $m = 1, 2$, the activities of the pre-output neurons $d_m(x, y, t)$ are given by

$$d_m\left(x, y, t\right) = \sum_k w_{m,k}(t) h_k(x, y, t),$$

in which $w_{m,k}(t)$ is the weight of the connection from hidden-layer neuron k to the pre-output neuron m for $m = 1, 2$ and $k = 1, \ldots, 900$. In this instance, $d_1\left(x, y, t\right)$ and $d_2\left(x, y, t\right)$ represent the positive and negative values, respectively. The final output of the network $v(x, y, t)$ integrated these values, as follows:

$$v\left(x, y, t\right) = d_1\left(x, y, t\right) - d_2\left(x, y, t\right).$$

During the interaction between the agent and the environment, the connection weights $w_{m,k}(t)$ were updated, depending on the prediction error, δ. To calculate δ, we used an action-value function $Q(s,a,t)$ for state s and action a at time t (Sutton and Barto, 2018). We defined $Q(s,a,t)$ by using the output of the network, as follows:

$$Q(s,a,t) = v\left(x',y',t\right),$$

in which $s = (x,y)$, $a = (\Delta x, \Delta y)$, $x^{'} = x + \Delta x$, and $y^{'} = y + \Delta y$. In our navigation tasks, $(\Delta x, \Delta y) = (1,0), (-1,0), (0,-1)$, or $(0,1)$, if the action was effective. The prediction error δ was represented by

$$\delta(t) = r(s,a) + \gamma Q\left(s^{'}, a^{'}, t\right) - Q(s,a,t),$$

in which $r(s,a)$ is the actual reward, $s^{'}$ denotes the state at time $t+1$, $a^{'}$ denotes the action at time $t+1$, and γ is the discount factor. We used temporal difference (TD) learning for the action-value function (i.e., state–action–reward–state–action [SARSA]) (Rummery and Niranjan, 1994) because the agent needed to remain at the same square after an ineffective action (i.e., hitting a wall). If we used classical TD learning for the state-value function, such ineffective actions would have disturbed the value learning. The connection weight $w_{1,k}(t)$ for $k = 1,\ldots,900$ was updated only when $\delta(t)$ was positive, and $w_{2,k}(t)$ for $k = 1,\ldots,900$ was updated only when $\delta(t)$ was negative. In the actual implementation, the updating rules were as follows:

$$w_{1,k}(t+1) = \begin{cases} w_{1,k}(t) + \frac{\alpha_1 \delta(t) h_k(x',y',t) d_1(x',y',t)}{N_1(x',y',t)}, \text{if } \delta(t) > 0 \\ w_{1,k}(t), \quad \text{otherwise,} \end{cases}$$

and

$$w_{2,k}(t+1) = \begin{cases} w_{2,k}(t) + \frac{\alpha_2 [-\delta(t)] h_k(x',y',t) d_2(x',y',t)}{N_2(x',y',t)}, \text{if } \delta(t) < 0 \\ w_{2,k}(t), \quad \text{otherwise,} \end{cases}$$

in which α_1 and α_2 are the learning rates and $N_1\left(x,y,t\right)$ and $N_2\left(x,y,t\right)$ are the normalization terms, which are represented as

$$N_m\left(x,y,t\right) = d_m\left(x,y,t\right) \sum_k h(x,y,t)^2$$

for $m = 1,2$. These normalization terms are introduced to make sure that.

$$Q(s,a,t+1) = \begin{cases} Q(s,a,t) + \alpha_1 \delta(t), \text{if } \delta(t) > 0 \\ Q(s,a,t) + \alpha_2 \delta(t), \text{otherwise.} \end{cases}$$

Algorithms for Comparison

We also implemented two representative algorithms to compare with the OVaRLAP model. One algorithm was simple TD learning (Sutton and Barto, 2018) (**Figure 1A**, right top); in our particular case, the algorithm was SARSA. In our implementation, a value function of states, $v_s\left(x,y,t\right)$, was represented as a look-up table and updated as

$$v_s\left(x,y,t+1\right) = v_s\left(x,y,t\right) + \alpha_s \delta_s(t)$$

in which α_s is the learning rate and δ_s is the prediction error, called the "TD error." This error is given by

$$\delta_s(t) = r(s,a) + \gamma_s Q_s\left(s',a',t\right) - Q_s(s,a,t)$$

in which γ_s is the discount factor. The action-value function, $Q_s(s,a,t)$, is given by

$$Q_s(s,a,t) = v_s\left(x',y',t\right).$$

The other algorithm is the MaxPain algorithm (Elfwing and Seymour, 2017) (**Figure 1A**, right bottom). This method is characterized by its distinct systems for the learning values for reward and pain (or punishment). The policy is dependent on the linear combination of the two value functions. In our implementation, we slightly modified the originally proposed MaxPain algorithm to make it comparable with the other methods, while maintaining the essential idea of MaxPain.

First, we used state-value functions of states $v_r\left(x,y,t\right)$ and $v_p\left(x,y,t\right)$, and their linear combination $v_L\left(x,y,t\right)$, to define the action-value functions $Q_r\left(s',a',t\right)$, $Q_p\left(s',a',t\right)$, and $Q_L\left(s^{'},a^{'},t\right)$, as follows:

$$Q_r(s,a,t) = v_r\left(x',y',t\right),$$

$$Q_p(s,a,t) = v_p\left(x',y',t\right),$$

and

$$Q_L(s,a,t) = v_L\left(x',y',t\right).$$

Second, we set the linear combination $v_L\left(x,y,t\right)$ without normalization, as follows:

$$v_L\left(x,y,t\right) = v_r\left(x,y,t\right) - v_p\left(x,y,t\right).$$

The couple of state-value functions were implemented as look-up tables and updated as

$$v_r\left(x,y,t+1\right) = v_r\left(x,y,t\right) + \alpha_r \delta_r(t)$$

and

$$v_p\left(x, y, t+1\right) = v_p\left(x, y, t\right) + \alpha_p \delta_p(t)$$

in which α_r and α_p are the learning rates for reward and pain, respectively, and δ_r and δ_p are the prediction errors for reward and pain, respectively. The prediction errors were calculated, depending on whether the reward observation was positive or negative, as follows:

$$\delta_r(t) = \varphi\left(r\left(s, a\right)\right) + \gamma_r Q_r\left(s^{'}, a^{'}, t\right) - Q_r\left(s, a, t\right)$$

and

$$\delta_p(t) = \varphi\left(-r\left(s, a\right)\right) + \gamma_p Q_p(s^{'}, \text{argmin}\left(Q_L\left(s^{'}, a^{'}, t\right)\right), t) \\ -Q_p\left(s, a, t\right)$$

in which $\varphi\left(z\right) = \max(z, 0)$ and γ_r and γ_p are the discount factors. A variant of off-policy Q-learning algorithm was used to calculate the pain prediction error δ_p, which enabled Q_p to learn for maximizing future pain (i.e., for predicting the worst case). These update rules were the same as those in the original study.

Action Selection

We used the softmax behavioral policy consistently for the three methods. It depends on the value function $\tilde{v}\left(x, y, t\right)$, i.e., $v\left(x, y, t\right)$ for OVaRLAP, $v_s\left(x, y, t\right)$ for the simple TD learning, and $v_L\left(x, y, t\right)$ for MaxPain. The probability that the agent selects an action a at position $\left(x, y\right)$ at time t is given by

$$\pi\left(a\middle|x, y, t\right) = \frac{\exp(\tilde{v}(x^{'}, y^{'}, t)/\tau)}{\sum_c \exp(\tilde{v}(x_c, y_c, t)/\tau)}$$

in which $x^{'}$ and $y^{'}$ denote the new state after selecting action a, x_c, and y_c denote the new state after selecting one of the possible actions, and τ is the temperature that controls the trade-off between exploration and exploitation. In our implementation, we used the common $\tau = 0.5$, for the three learning methods.

Painful Grid-World Navigation Task

The purpose of this task was to navigate from the starting position to either of the two goals, while avoiding hitting the wall. Possible actions at each time step were moving one step north, south, east, and west. For example, if the agent moved one step north, $(\Delta x, \Delta y) = (0, 1)$. Two goals exist with reward of 1 or 2. If the agent hits a wall, then it received a negative reward of -1 and remained at the same position. This was an exceptional case. In other cases, the agent could by necessity move to the next square. An episode began when the agent started from the starting position and ended when the agent reached either of the goals. The agent repeated such learning episodes.

A single run consisted of 500 learning episodes, after initializing the value function so that the value for each state was zero. We ran 50 separate runs each for OVaRLAP, simple TD learning, and MaxPain with a single grid-world configuration. We conducted five simulation experiments for each of which we used a grid-world with a consistent character but different configuration. The structure of each grid-world is shown in **Figure 1B** and in **Supplementary Figure S1**. We tested OVaRLAP with various amount of generalization by setting θ as $\theta \in \{0.44, 0.66, 1.0, 1.5, 2.2\}$. The value of θ was fixed in each simulation experiment. We set $A = \rho = 0$, that is, no noise in the fixed connections in OVaRLAP. We set the other metaparameters for each algorithm as follows: $\alpha_1 = \alpha_2 = 0.1$, and $\gamma = 0.95$ for OVaRLAP; $\alpha_s = 0.1$ and $\gamma_s = 0.95$ for simple TD learning; and $\alpha_r = \alpha_p = 0.1$, $\gamma_r = 0.95$, and $\gamma_p = 0.5$ for MaxPain. OVaRLAP and MaxPain are extended algorithms from simple TD; therefore, they used common metaparameters with simple TD, when they could share them.

The five configurations in **Supplementary Figure S1** were generated, based on the following rules, while avoiding symmetric or similar structures. Each configuration had a wide passage from which a narrow passage branched off. These passages had fixed widths and lengths. The starting position and the goal with a reward of one were at either end of a wide passage. A goal with reward of two was at the end of a narrow passage. The legend for **Supplementary Figure S1** provides further details of the rules.

Grid-World Navigation Task for Disturbed OVaRLAP

The purpose of this task was to navigate from the starting position to either of four goals. The structure of the grid-world is shown in **Figure 1D**. Possible actions at each time step were moving one step north, south, east, and west. Multiple goals existed, each of which gave a reward of one. Our interest in this experiment was not in pain aversion; therefore, we did not apply a negative reward to the actions that resulted in hitting a wall (i.e., no pain). The agent simply remained at the same position after hitting a wall. Even in this no pain setting, the agents after reinforcement learning likely avoid wall hits, due to the discount factor in the value functions. An episode began when the agent moved from the starting position and ended when the agent reached one of the four goals. A single run consisted of 40,000 time steps in total (i.e., 378.8 learning episodes on average) after the value function initialization at the onset of the first learning episode. For each metaparameter setting described below, we performed 50 separate runs for OVaRLAP. We considered the following cases: normal or no update for a negative TD error (i.e., "intact" or "impaired") and with or without noise in the fixed connections (i.e., "noised" or "unnoised"). There were 25 different settings for the noised case, consisting of pairs of noise strength A and noise fraction ρ, taken from $A \in \{0.25, 0.5, 1, 2, 4\}$ and $\rho \in \{0.00125, 0.0025, 0.005, 0.01, 0.02\}$, respectively. Although the values of A and ρ were fixed within each simulation setting, the noise, $\varepsilon_{ij,k}$, was generated independently for each of the 50 runs and was fixed within each run. We set the other metaparameters, as follows: $\theta = 1$; $A = \rho = 0$ for the unnoised case; $\alpha_1 = 0.1$; $\alpha_2 = 0.1$ for the intact setting (which was the same as in the previous experiment) and $\alpha_2 = 0$ for the impaired setting; $\gamma = 0.8$; and $\tau = 0.5$.

RESULTS

The Behavior of OVaRLAP

We evaluated OVaRLAP in a spatial navigation task in painful grid-worlds by comparing its performance with those of simple TD learning and MaxPain (Elfwing and Seymour, 2017). **Figure 1A** shows a schematic representation of these algorithms. OVaRLAP achieves stimulus generalization based on the similarity of states (**Figure 2A**) because it utilizes a fixed neural network (gray arrows in **Figure 1A**, left) for the preprocessing of the value learning network (blue and red arrows in **Figure 1A**, left). By contrast, the other two algorithms, which have look-up tables, show no stimulus generalization (**Figures 2B,C**). The task in this study was to seek rewarded goals while avoiding painful wall hits in which the agent had to manage the trade-off between obtaining rewards and avoiding pain (**Figure 1B**). We used five different grid-worlds, each of which had a safe but lowly rewarded goal and a risky but highly rewarded goal (**Supplementary Figure S1**).

Figures 3A–C show the average learning curves at each episode over 50 independent runs for the grid-world shown in **Figure 1B**. The reward per step of the OVaRLAP agent was higher than that of the MaxPain agent, but it did not exceed that of the simple TD learning agent (**Figure 3A**). Compared to the other agents, the OVaRLAP agent reached each of the goals with a smaller number of steps in the early learning phase (**Figure 3B**). After 100 learning episodes, all three agents showed a similar number of steps to reach either of the goals. With regard to pain aversion, the OVaRLAP agent also showed quick learning. The OVaRLAP agent exhibited fewer wall hits in the early learning phase than the other agents did. However, after 50 learning episodes, the number of wall hits of the OVaRLAP agent was comparable to that of the simple TD learning agent (**Figure 3C**). By contrast, the MaxPain agent first hit the walls as many times as the simple TD learning agent, but the number of wall hits then quickly decreased. The OVaRLAP agent showed its characteristic performance in all of the five simulation experiments, each of which had a different grid-world with a similar spatial structure (**Figures 3D–F**). The relative performance of the three agents was consistent when the starting position was randomized (**Supplementary Figure S2**). We also tested the OVaRLAP agents with various levels of generalization (**Figure 4A**). The results showed that an increase in generalization led to better performance unless the generalization became too large, whereas too large generalization deteriorated the performance (**Figure 4B**). These results suggest that the OVaRLAP agent could learn a reward approach and pain aversion in a very efficient manner owing to its proper generalization, whereas it was inferior to the simple TD in long-term reward learning and to MaxPain in long-term pain aversion.

Figure 5 shows how the value function developed using the three methods, in which OVaRLAP exhibited unique profiles (**Figure 5**, top). First, the values of positions close to the walls decreased, as did the values of the positions of the walls. After only five learning episodes, this generalization constructed a safe passage to the low-reward goal while simultaneously causing a dip in the value function as a hazard to approach the risky but high-reward goal. This value hazard completely disappeared after 500 episodes because the reward learning had propagated to the squares along the passage to the high-reward goal. Thus, quick pain learning was essential for aggressive pain aversion and conservative reward-seeking in the early learning stage.

The simple TD learning algorithm did not construct any value hazard to the high-reward goal during its learning process because no special system existed for pain learning (**Figure 5**, middle). The MaxPain algorithm also produced hazards on the passage to the high-reward goal (**Figure 5**, bottom). In contrast to OVaRLAP, the hazard progressively increased as the learning proceeded and never flattened owing to its strong pain learning ability. For this reason, the MaxPain algorithm persistently maintained low values on the way to the high-reward goal.

We further evaluated how the model agent navigates if the environment is changed to have pain-free walls after the agent has been trained with painful walls (other than walls, everything is the same before and after the change). Interestingly, the OVaRLAP agent quickly unlearnt the past painful stimuli, compared to the other two models (**Figure 6**). This result indicates that generalization worked not only in the early phase of learning but also contributed to relearning timings induced by environmental changes.

The Effect of a Disturbed OVaRLAP

We next examined how the learning behaviors of OVaRLAP were disturbed by disabling weight updating by negative TD errors and introducing noise to the fixed network (**Figure 1C**). The generalization after obtaining a reward became a little noisy because of the noise introduction (**Figure 7A**). We applied this disturbed OVaRLAP to a grid-world navigation task without pain (**Figure 1D**). We tested each of the following two-by-two settings: normal or no update for the negative TD error (i.e., "intact" or "impaired"), and with or without noise in the fixed network (i.e., "noised" or "unnoised").

The impaired noised agent acquired an aberrant value function: it increased the value for a specific position distant from the goals, as if the position would give a positive reward (**Figure 7B**, bottom right). The value function of the intact noised agent was slightly noisy (**Figure 7B**, top right), but nearly the same as that of the intact unnoised agent (**Figure 7B**, top left). The impaired unnoised agent made the values around the goals higher than those of the intact agents, but the values for positions distant from the goals remained low (**Figure 7B**, bottom left).

We ran 50 runs for each setting to confirm reproducibility. For the noised agent, we applied a different pattern of noise in each run. **Figure 7C** presents the maximum value and its position in the value function after 40,000 steps in each run. The intact agents always established the maximum values at the positions of goals and maintained them consistent to the actual amount of the rewards the agent

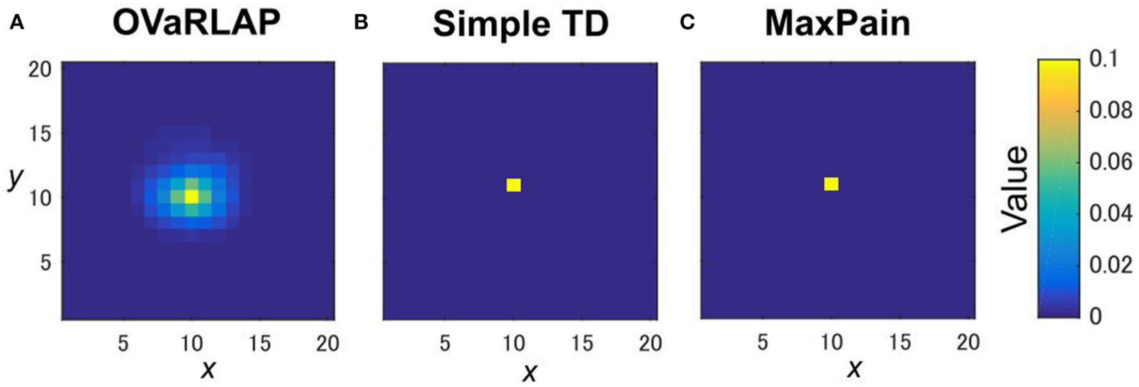

FIGURE 2 | State values after a single update. The value function is updated after receiving a positive reward of one once with OVaRLAP **(A)**, simple TD learning **(B)**, and MaxPain **(C)** at the center of a two-dimensional grid-world in **Figure 1B**. For OVaRLAP, the metaparameter θ is set as $\theta = 1$ to introduce a moderate amount of generalization.

obtained (**Figure 7C**, top row). The maximum values of the impaired unnoised agent were higher than the actual amount of the rewards (**Figure 7C**, bottom left). Their positions were not necessarily the same as the actual goals, but they were, at most, two steps away from the goals. In contrast to these agents, the impaired noised agent produced maximum values that were quite apart from the actual goals (**Figure 7C**, bottom right). These values were higher than the actual amount of rewards and were sometimes exceedingly high.

For the noised case, we tested different noise settings by varying the noise strength and the noise fraction (**Supplementary Figure S3**), and for each we performed 50 runs. **Supplementary Table S1** provides a summary of the results and shows that the various levels of noise induced aberrant valuation if the negative value neuron was impaired. It also shows that sole large noise did not lead to aberrant valuation as long as the negative value learning was intact.

DISCUSSION

The Effect of Generalization and Discrimination on Behaviors

In this study, we proposed a novel reinforcement learning model named OVaRLAP to analyze how the computational characteristics in the striatal dopamine system affect behaviors. The OVaRLAP model reproduced stimulus generalization for both reward learning and punishment learning. Discrimination followed generalization to shape the value function, that is, inhibiting excessive expectation of reward or punishment caused by generalization. It was realized by the activity of one pre-output neuron that offset the activity of the other pre-output neuron. In the navigation task in painful grid-worlds (**Figures 3–5**), punishment learning due to painful wall hits was first generalized and subsequent unexpected safeness caused discrimination. By contrast, in the case of painless grid-world (**Figure 7**), generalization for reward learning was followed by discrimination due to the unexpected unrewarded outcome.

The OVaRLAP model enabled safe and efficient exploration in the painful grid-world navigation task. The transition of the value function for the OVaRLAP agent shows how stimulus generalization and discrimination contribute to managing the trade-off between reward-seeking and pain aversion (**Figure 5**, top).

First, punishment learning due to hitting walls generalized quickly, as shown in the value function after five episodes. It led to a preference for the center of the passage (i.e., pain aversion). Indeed, the OVaRLAP agent first showed fewer wall hits than the simple TD learning and MaxPain agents did (**Figures 3C,F**). This contributed to reducing the number of steps to reach either of the goals in the early learning phase (**Figures 3B,E**). Thereafter, The OVaRLAP agent increased the values of positions close to the walls. This transition indicated that the agent discriminated between the safe and painful positions. This discrimination increased the tendency to reach the high-reward goal. In short, stimulus generalization of punishment learning induced pain aversion, followed by discrimination for switching to reward-seeking. The simple TD learning agent did not show pain aversion (**Figure 5**, middle). Its value function was optimized to maximize future reward, which is consistent with its high reward per step (**Figures 3A,D**). The MaxPain agent showed strong pain aversion based on its separate value learning system to expect future pain (**Figure 5**, bottom row). It achieved very few wall hits (**Figures 3B,E**) while maintaining pain aversion and low values on the way to the high-reward goal, and this corresponded to a lower reward per step compared to the other agents (**Figures 3A,D**).

Stimulus generalization and discrimination were effective in safe and efficient exploration; however, dysfunction in

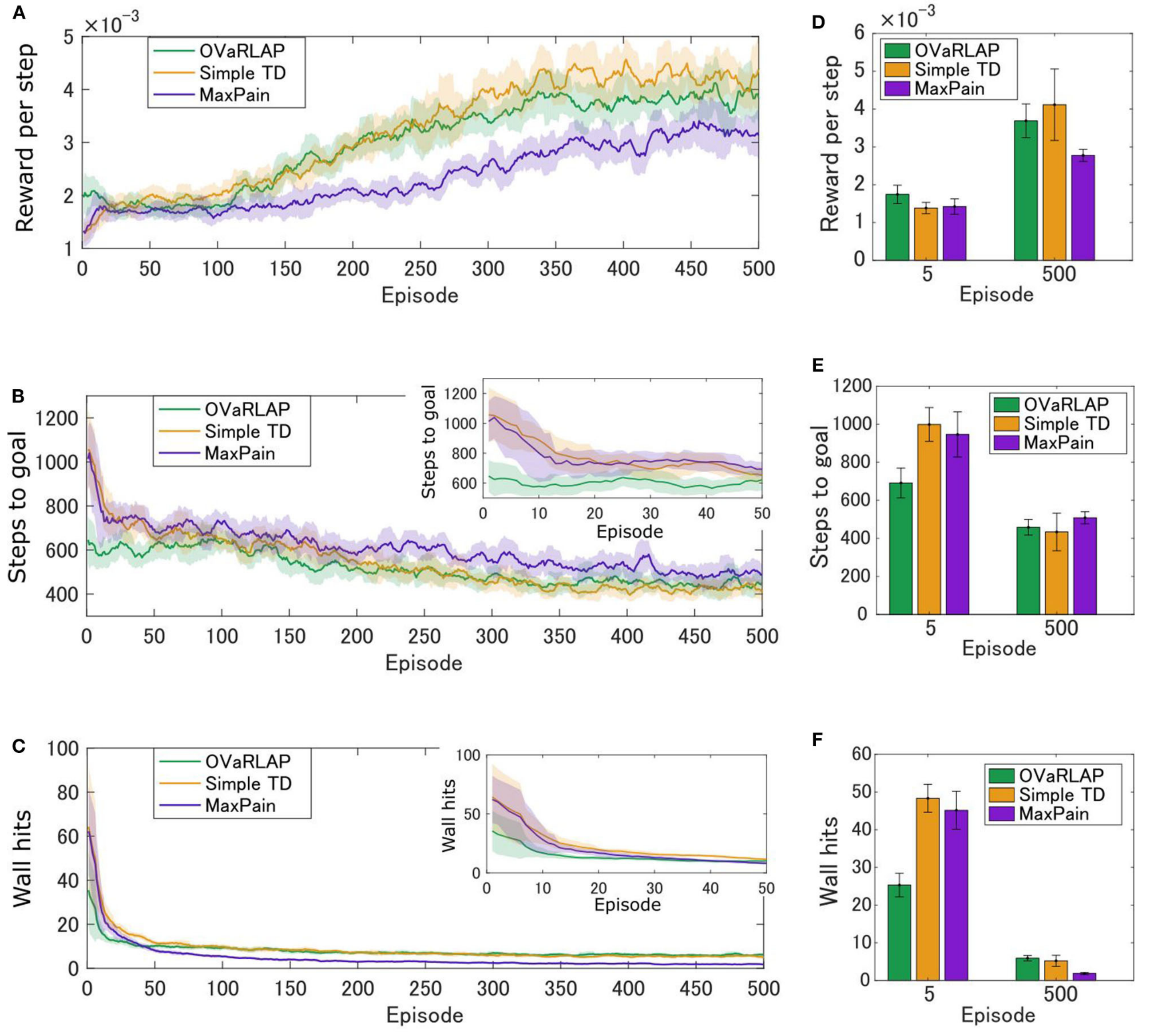

FIGURE 3 | Average performance of OVaRLAP, the simple TD learning, and MaxPain in the painful grid-world navigation task. **(A)** The average reward per step, **(B)** the average number of steps to either of the two goals, and **(C)** the average number of wall hits, based on the three types of agents after each number of learning episodes on the horizontal axis. The grid world shown in **Figure 1B** was used. The number of steps to either of the two goals and the number of wall hits for the first 50 learning episodes are magnified in the top right insets of **(B,C)**, respectively. Each average (on the vertical axis) was obtained over 50 separate runs and then smoothed using a moving average over the preceding 11 episodes (on the horizontal axis). The thick lines represent the moving averages and the shadow areas indicate the moving standard deviations. **(D–F)** The average performance over five simulation experiments, each of which used a grid-world with different configurations. The grid world shown in **Figure 1B** and its variants were used (**Supplementary Figure S1** presents the details of the variants). **(D)** The average reward per step, **(E)** the average number of steps to either of the two goals, and **(F)** the average number of wall hits, based on the three types of agents after five learning episodes and 500 learning episodes, are shown. Each bar indicates the average over five grid-world configurations, after taking the average over 50 separate runs with each grid-world configuration. Each error bar represents the standard deviation over the five configurations. For OVaRLAP, the metaparameter θ was set as $\theta = 1$ to induce a moderate amount of generalization.

the system may induce aberrant learning behaviors. First, too large of a generalization deteriorated the performance of the OVaRLAP agent (**Figure 4**), which is consistent with the previous reports that generalization to an abnormal extent is associated with psychiatric disorders (Dunsmoor and Paz, 2015; Asok et al., 2018). Second, the impaired update for

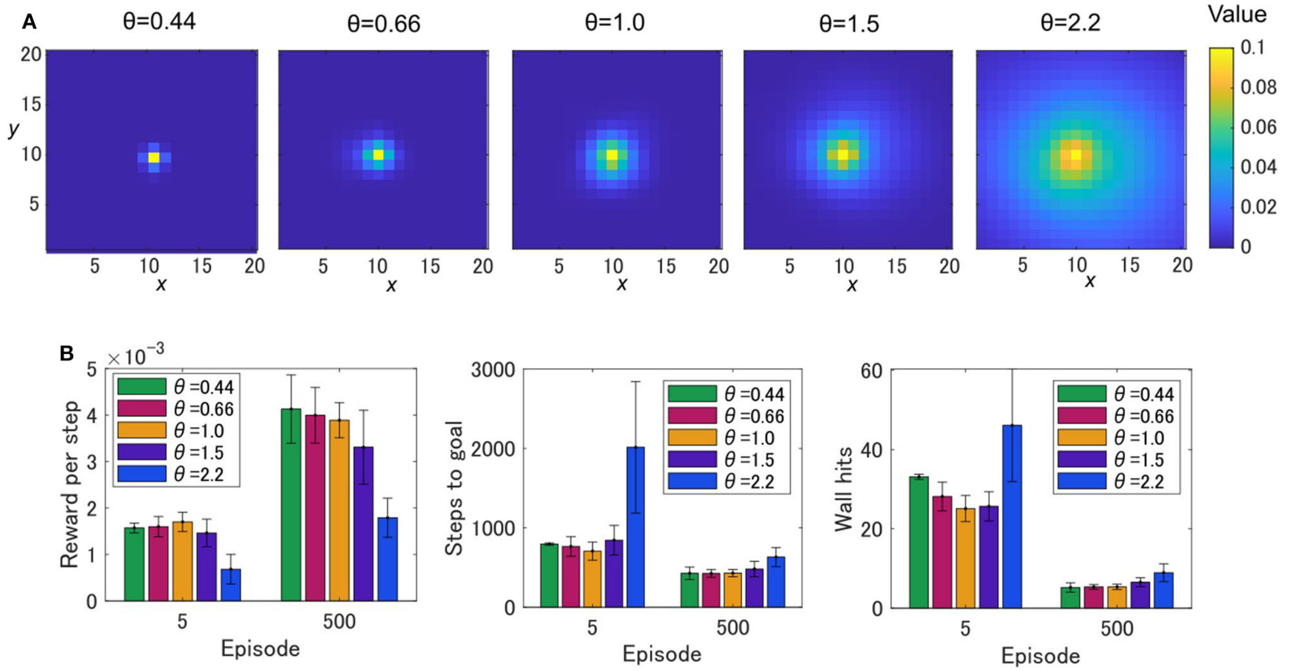

FIGURE 4 | The painful grid-world navigation task for OVaRLAP with various amount of generalization. **(A)** State values after a single update for OVaRLAP with a variety of metaparameter settings, $\theta \in \{0.44,\ 0.66,\ 1.0,\ 1.5,\ 2.2\}$, which control the level of generalization. The value function is updated after receiving a positive reward of one once at the center of a two-dimensional grid-world in **Figure 1B**. **(B)** The average performance over five simulation experiments, each of which used a grid-world with a different configuration than the others. The grid-world shown in **Figure 1B** and its variants were used (**Supplementary Figure S1** presents the details of the variants). The average reward per step (left), the average number of steps to either of the two goals (middle), and the average number of wall hits (right), over five OVaRLAP agents with different metaparameter settings, $\theta \in \{0.44,\ 0.66,\ 1.0,\ 1.5,\ 2.2\}$, after five learning episodes and 500 learning episodes, are shown. Each bar indicates the average over five grid-world configurations, after taking the average over 50 separate runs with each grid-world configuration. Each error bar represents the standard deviation over the five configurations.

the negative TD error combined with the noise in the fixed connections induced aberrant valuation in OVaRLAP (**Figure 7**, **Supplementary Table S1**). The value increased by noisy stimulus generalization was not reshaped by discrimination because of the impairment of the punishment learning. A single update of the value was only slightly affected by the noisy stimulus generalization (**Figure 7A**); however, as the number of arrivals to the actual goals increased, such aberrant updates of the value function would have allotted abnormally high values to some positions that were actually of no reward (**Figures 7B–C**).

Computation Underlying Generalization

The structure of OVaRLAP provides insight into the neurobiological basis of stimulus generalization and discrimination. The two pre-output neurons separately responded to positive and negative prediction errors to update the connections between the hidden-layer neurons and the corresponding pre-output neuron. This hybrid learning system exhibits good correspondence to the striatal dopamine system in which D1- and D2-SPNs differentially respond to dopamine so that the connections between the cortex and the striatal SPNs are differently modulated (Reynolds et al., 2001; Hikida et al., 2013; Yagishita et al., 2014). Transforming the input into the hidden-layer activity approximates the process by which an external stimulus and the internal state are encoded into neural activity patterns of the cortex. This process is assumed to be rather independent from dopamine-dependent plasticity. However, this process should not be based simply on random connections because the similarity of the inputs has to be maintained in the encoding process to achieve stimulus generalization. Instead, encoding could be learned in a different manner from the reward-related learning. Therefore, the fixed connections between the input and hidden layers in OVaRLAP can be the result of learning for such encoding process. How the brain learns the encoding process is a topic for future research, but some implications are derived from computational models of the primary visual cortex and the primary auditory cortex that adopted unsupervised learning (Hyvärinen and Hoyer, 2001; Terashima and Okada, 2012).

The encoding process may also depend on additional functions other than learning. It has been reported that the prefrontal dopamine system could be involved in processing incoming sensory signals such as working memory and attention (Ott and Nieder, 2019). Although we have not modeled the prefrontal dopamine system in the current

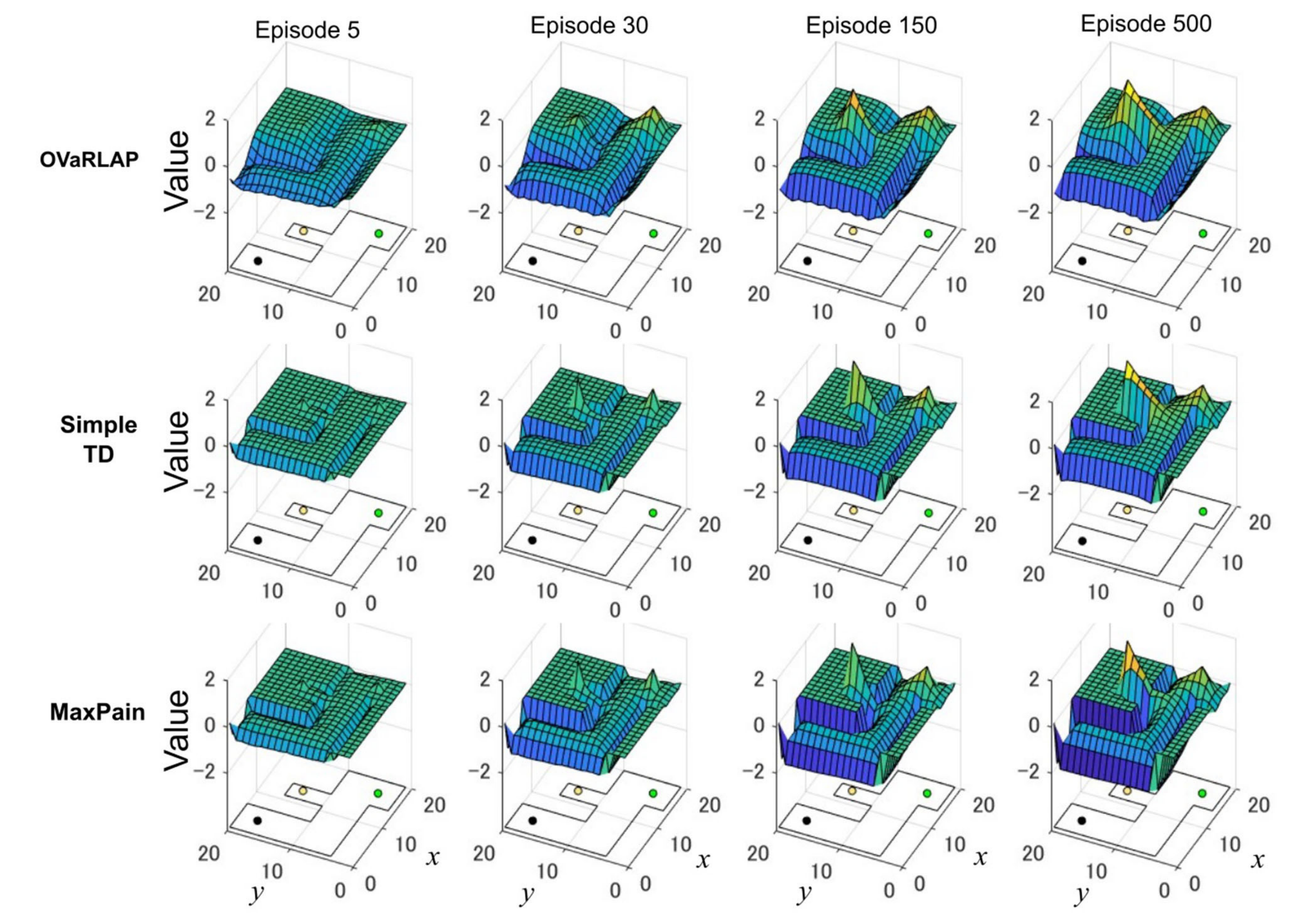

FIGURE 5 | Transition of the value function for each model in the grid-world navigation task. The average state values over 50 separate runs after 5, 30, 150, and 500 learning episodes for OVaRLAP (top row), simple TD learning (middle row), and MaxPain (bottom row), when applied to the painful grid-world navigation task shown in **Figure 1B**. An outline of the grid-world is shown on the floor of each plot. The black lines indicate walls. The black, green, and yellow circles are the starting position, the low-reward goal, and the high-reward goal, respectively. For OVaRLAP, the metaparameter θ is set as $\theta = 1$ to introduce a moderate level of generalization. For MaxPain, the result of the subtraction of the value function for the pain from the value function for the positive reward (i.e., the main factor for decision-making) is displayed.

OVaRLAP model, one possible implementation would allow prefrontal dopamine to modify the information transmission from the input layer to the hidden layer and thereby alter generalization.

Compared to ordinary reinforcement learning algorithms, the uniqueness of the OVaRLAP model is that it is characterized by fast generalization and separation of positive and negative value learning. Its advantages in safe and efficient exploration primarily appeared due to fast generalization, whereas its potential defect in causing maladaptive behaviors was attributed to fast generalization and also separation of positive and negative value learning. Mikhael and Bogacz previously demonstrated the advantage of separation of positive and negative values (Mikhael and Bogacz, 2016). Their model coded reward uncertainty into the sum of synaptic weights of D1- and D2-SPNs. In addition, their model could adjust for the tendency to choose or avoid risky options by changing the weight of the positive and negative values, which is assumed to reflect tonic dopamine levels. This model is consistent with risk-taking behaviors observed in patients of Huntington's disease, in which striatal neurons are closely involved (Kalkhoven et al., 2014). To further discuss advantages of separation of positive and negative values is beyond the scope of our current study, although the OVaRLAP model may also exhibit such advantages. The OVaRLAP model can encode reward uncertainty in distinct connections updated by positive and negative TD errors (red and blue arrows, respectively, in **Figure 1A**, left), and can adjust a risk-taking tendency by changing the connection weights between the pre-output neurons to the output neuron, which were set as constants for simplicity in the current study.

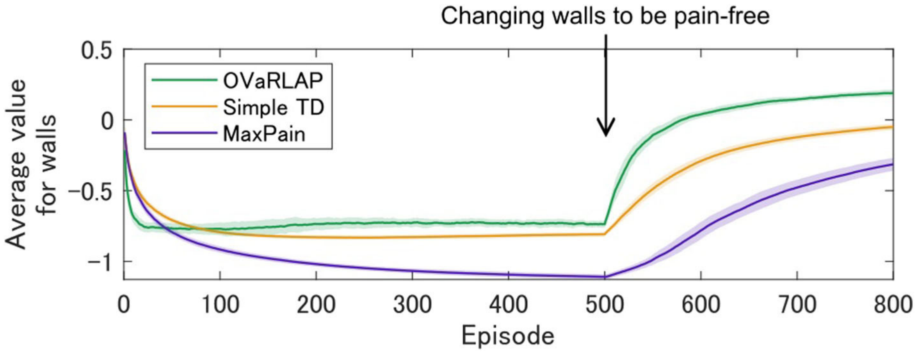

FIGURE 6 | Relearning values of walls after environmental changes. OVaRLAP, the simple TD learning, and MaxPain were applied to a grid-world navigation task in which the agent was trained with painful walls until the 500th episode and the environment was changed at the 500th episode to have pain-free walls (other than walls, everything was the same before and after the change). The grid-world shown in **Figure 1B** was used. The state-value averaged over all wall positions is plotted for each of the three agents, on the horizontal axis indicating the number of learning episodes. Each average (on the vertical axis) is obtained over 50 separate runs and the shadow areas indicate the standard deviation. For OVaRLAP, the metaparameter θ is set as $\theta = 1$ to introduce a moderate level of generalization.

Hypotheses on Psychiatric Disorders

The potential defect of OVaRLAP that causes maladaptive behaviors could provide a hypothesis on psychiatric disorders. Based on the correspondence between OVaRLAP and the brain, our results with the disturbed OVaRLAP showed the possibility that the noisy encoding process in the cortex and the impairment of dopamine-dependent plasticity induce abnormal stimulus generalization and discrimination, which may underlie delusional symptoms. The relationship between cortical disconnectivity and disrupted learning in schizophrenia has been implied in computational studies (Hoffman and Dobscha, 1989; Hoffman and McGlashan, 1997); therefore, alterations in cortical connectivity could be a cause of a noisy encoding process. Histological examinations and diffusion tensor imaging studies of patients with schizophrenia revealed reduced dendritic spine density and disrupted white matter connectivity, respectively (Garey et al., 1998; Ellison-Wright and Bullmore, 2009; van den Heuvel et al., 2016). Reinforcement learning models have been used for attempts to explain positive symptoms, including delusions, in schizophrenia (Deserno et al., 2013; Katahira and Yamashita, 2017; Maia and Frank, 2017). However, these studies do not link abnormal cortical connectivity to positive symptoms. Rather, they attribute positive symptoms to aberrant salience (i.e., a surprise response to non-salient events) (Kapur, 2003). Thus, they focused on abnormalities in the dopamine system, which was also supported by neurobiological evidence (Howes et al., 2012; Daberkow et al., 2013). The OVaRLAP model is a new corticostriatal learning model that relates delusional symptoms to abnormalities in cortical connectivity and the dopamine system.

The relationship between abnormal stimulus generalization and schizophrenia has been investigated in psychological research (Buss and Daniell, 1967; Ralph, 1968; Kahnt and Tobler, 2016) based on the hypothesis that abnormal generalization underlies delusional symptoms in schizophrenia. Some studies imply heightened stimulus generalization in schizophrenia (Ralph, 1968; Kahnt and Tobler, 2016). However, the reported abnormality was not sufficiently remarkable for explaining delusion in a straightforward manner. This could be attributed to the conditioning paradigm used for their behavioral experiments. Patients with schizophrenia have various deficits in cognitive function; therefore, the experimental design needs to be simple to attribute the result to stimulus generalization rather than other factors. Computational models may potentially be used to investigate how a specific deficit in cognitive function leads to psychotic symptoms through complex learning processes. The OVaRLAP model showed that a small abnormality in stimulus generalization in combination with an unbalanced response to positive and negative prediction error may induce aberrant valuation after reward-related learning, including action selection, state transition, and obtaining rewards from multiple sources.

Conclusions

The OVaRLAP model updates synaptic weights of the last layer in a multilayer neural network, which reflects dopamine-dependent plasticity of corticostriatal synapses (Reynolds et al., 2001). In contrast to ELM (Huang et al., 2006), where fixed connections between the input and hidden layers are set randomly, we set the fixed connections of the OVaRLAP model to maintain the similarity of inputs in the hidden-layer representation. The OVaRLAP model enabled fast generalization of reward and punishment learning. In the painful grid-world navigation tasks, it demonstrated a quick reward approach and efficient pain aversion in the early learning phase and achieved safe and efficient exploration. However, disturbances of the OVaRLAP network that caused

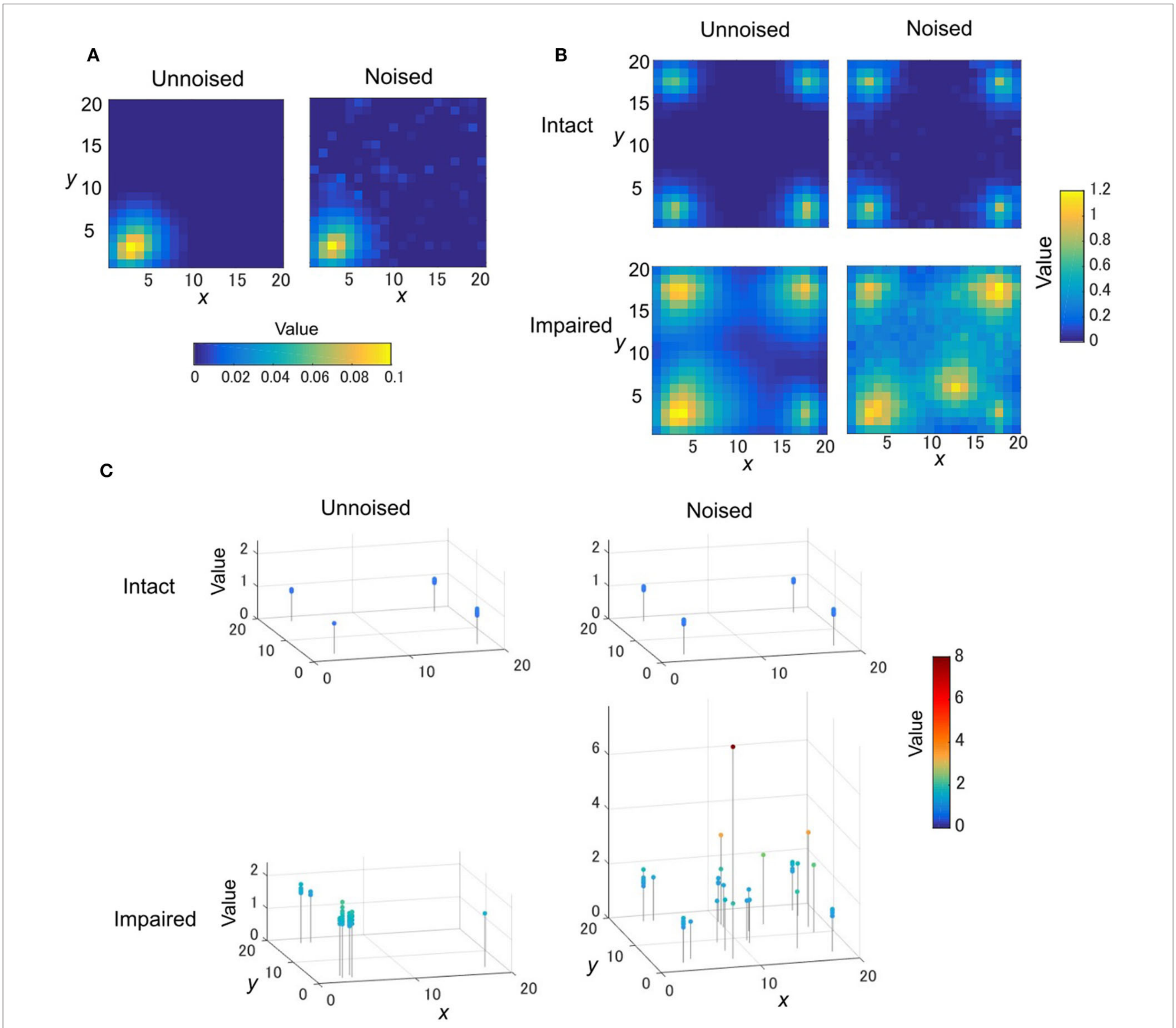

FIGURE 7 | The TD learning task for disturbed OVaRLAP. **(A)** The value function is updated after reaching the left-bottom goal of the grid-world shown in **Figure 1D** once with OVaRLAP without noise (left) and with noise (right). **(B)** The value function after 120 learning episodes in a single run for OVaRLAP with intact and impaired updates (top and bottom rows, respectively) for the negative TD error with and without noise (right and left columns, respectively) in the fixed connections. **(C)** The maximum value and its position in the value function after 40,000 time steps in total (378.8 learning episodes on average) in each run. Each panel shows the maximum values in 50 separate runs for OVaRLAP under the following settings: intact or impaired updates (top or bottom row, respectively) for the negative TD error, and without noise or with noise (left or right column, respectively) in the fixed connections. In the noisy case, the noise was independently introduced to the initialization of the fixed connections for each run.

noisy generalization and impaired discrimination led to aberrant valuation.

These results suggested the advantage and potential drawback of generalization by the striatal dopamine system with regard to adaptive behaviors. These results are consistent with previous theories in behavioral science (Dunsmoor and Paz, 2015; Kahnt and Tobler, 2016; Asok et al., 2018), in which generalization is considered to be adaptive, whereas abnormal generalization is implicated in maladaptive behaviors. The OVaRLAP model also gives insight into the neurobiological basis of generalization and its dysfunction. The process for encoding external stimuli and internal states into neural activity patterns of the cortex may be learned independently from reward-related learning. Disruption of this encoding process induced by altered cortical connectivity may disturb reward- and punishment-related learning, possibly underlying delusional symptoms of psychiatric disorders.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

YF, SY, HK, and SI conceived the study. YF developed, analyzed, and simulated the models. YF and SI wrote the paper. SY and HK reviewed drafts of the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by Grants-in-Aids for Scientific Research on Innovative Areas (17H06310 to YF and SI), Science Research (B) (19H04180 to SI), the Program for Brain/MINDS (to SI) from AMED, CREST (JPMJCR1652 to HK and SI) from JST, and World Premier International Research Center Initiative support (to SY, HK, and SI) from JSPS.

ACKNOWLEDGMENTS

We would like to thank Hidetoshi Urakubo for his helpful comments. This manuscript has been released as a pre-print at https://doi.org/10.1101/2019.12.12.873950 (Fujita et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00066/full#supplementary-material

REFERENCES

Amemori, K., Gibb, L. G., and Graybiel, A. M. (2011). Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments. *Front. Hum. Neurosci*. 5:47. doi: 10.3389/fnhum.2011.00047

Asok, A., Kandel, E. R., and Rayman, J. B. (2018). the neurobiology of fear generalization. *Front. Behav. Neurosci*. 12:329. doi: 10.3389/fnbeh.2018.00329

Bordi, F., and LeDoux, J. (1992). Sensory tuning beyond the sensory system: an initial analysis of auditory response properties of neurons in the lateral amygdaloid nucleus and overlying areas of the striatum. *J. Neurosci.* 12, 2493–2503. doi: 10.1523/JNEUROSCI.12-07-02493.1992

Buss, A. H., and Daniell, E. F. (1967). Stimulus generalization and schizophrenia. *J. Abnorm. Psychol*. 72, 50–53. doi: 10.1037/h0020082

Collins, A. G., and Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev*. 121, 337–366. doi: 10.1037/a0037015

Cox, J., and Witten, I. B. (2019). Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* 20, 482–494. doi: 10.1038/s41583-019-0189-2

Daberkow, D. P., Brown, H. D., Bunner, K. D., Kraniotis, S. A., Doellman, M. A., Ragozzino, M. E., et al. (2013). Amphetamine paradoxically augments exocytotic dopamine release and phasic dopamine signals. *J. Neurosci.* 33, 452–463. doi: 10.1523/JNEUROSCI.2136-12.2013

DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends Neurosci*. 13, 281–285. doi: 10.1016/0166-2236(90)90110-V

Deserno, L., Boehme, R., Heinz, A., and Schlagenhauf, F. (2013). Reinforcement learning and dopamine in schizophrenia: dimensions of symptoms or specific features of a disease group? *Front. Psychiatry* 4:172. doi: 10.3389/fpsyt.2013.00172

Doya, K. (2007). Reinforcement learning: computational theory and biological mechanisms. *HFSP J*. 1, 30–40. doi: 10.2976/1.2732246/10.2976/1

Dunsmoor, J. E., and Paz, R. (2015). Fear generalization and anxiety: behavioral and neural mechanisms. *Biol. Psychiatry* 78, 336–343. doi: 10.1016/j.biopsych.2015.04.010

Elfwing, S., and Seymour, B. (2017). "Parallel reward and punishment control in humans and robots: safe reinforcement learning using the MaxPain algorithm," in *Paper Presented at the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Lisbon).

Ellison-Wright, I., and Bullmore, E. (2009). Meta-analysis of diffusion tensor imaging studies in schizophrenia. *Schizophr. Res*. 108, 3–10. doi: 10.1016/j.schres.2008.11.021

Frank, M. J., Seeberger, L. C., and O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943. doi: 10.1126/science.1102941

Franks, D. W., and Ruxton, G. D. (2008). How robust are neural network models of stimulus generalization? *BioSystems* 92, 175–181. doi: 10.1016/j.biosystems.2008.02.003

Fujita, Y., Yagishita, S., Kasai, H., and Ishii, S. (2019). Computational characteristics of the striatal dopamine system described by reinforcement learning with fast generalization. *bioRxiv [Preprint]*. doi: 10.1101/2019.12.12.873950

Garey, L. J., Ong, W. Y., Patel, T. S., Kanani, M., Davis, A., Mortimer, A. M., et al. (1998). Reduced dendritic spine density on cerebral cortical pyramidal neurons in schizophrenia. *J. Neurol. Neurosurg. Psychiatry* 65, 446–453. doi: 10.1136/jnnp.65.4.446

Ghirlanda, S., and Enquist, M. (1998). Artificial neural networks as models of stimulus control. *Anim. Behav*. 56, 1383–1389. doi: 10.1006/anbe.1998.0903

Ghirlanda, S., and Enquist, M. (2003). A century of generalization. *Anim. Behav*. 66, 15–36. doi: 10.1006/anbe.2003.2174

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 3), 15647–15654. doi: 10.1073/pnas.1014269108

Hikida, T., Yawata, S., Yamaguchi, T., Danjo, T., Sasaoka, T., Wang, Y., et al. (2013). Pathway-specific modulation of nucleus accumbens in reward and aversive behavior via selective transmitter receptors. *Proc. Natl. Acad. Sci. U.S.A*. 110, 342–347. doi: 10.1073/pnas.1220358110

Hoffman, R. E., and Dobscha, S. K. (1989). Cortical pruning and the development of schizophrenia: a computer model. *Schizophr. Bull*. 15, 477–490. doi: 10.1093/schbul/15.3.477

Hoffman, R. E., and McGlashan, T. H. (1997). Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "voices" in schizophrenia. *Am. J. Psychiatry* 154, 1683–1689. doi: 10.1176/ajp.154.12.1683

Howes, O. D., Kambeitz, J., Kim, E., Stahl, D., Slifstein, M., Abi-Dargham, A., et al. (2012). The nature of dopamine dysfunction in schizophrenia and what this means for treatment. *Arch. Gen. Psychiatry* 69, 776–786. doi: 10.1001/archgenpsychiatry.2012.169

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Hyvärinen, A., and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res*. 41, 2413–2423. doi: 10.1016/S0042-6989(01)00114-6

Iino, Y., Sawada, T., Yamaguchi, K., Tajiri, M., Ishii, S., Kasai, H., et al. (2020). Dopamine D2 receptors in discrimination learning and spine enlargement. *Nature* 579, 555–560. doi: 10.1038/s41586-020-2115-1

Kahnt, T., and Tobler, P. N. (2016). Dopamine regulates stimulus generalization in the human hippocampus. *Elife* 5:e12678. doi: 10.7554/eLife.12678

Kalkhoven, C., Sennef, C., Peeters, A., and van den Bos, R. (2014). Risk-taking and pathological gambling behavior in Huntington's disease. *Front. Behav. Neurosci.* 8:103. doi: 10.3389/fnbeh.2014.00103

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am. J. Psychiatry* 160, 13–23. doi: 10.1176/appi.ajp.160.1.13

Katahira, K., and Yamashita, Y. (2017). A theoretical framework for evaluating psychiatric research strategies. *Comput. Psychiatry* 1, 184–207. doi: 10.1162/CPSY_a_00008

Lukoševičius, M., and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* 3, 127–149. doi: 10.1016/j.cosrev.2009.03.005

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput*. 14, 2531–2560. doi: 10.1162/089976602760407955

Maia, T. V., and Frank, M. J. (2017). An integrative perspective on the role of dopamine in schizophrenia. *Biol. Psychiatry* 81, 52–66. doi: 10.1016/j.biopsych.2016.05.021

Meredith, G. E., Baldo, B. A., Andrezjewski, M. E., and Kelley, A. E. (2008). The structural basis for mapping behavior onto the ventral striatum and its subdivisions. *Brain Struct. Funct.* 213, 17–27. doi: 10.1007/s00429-008-0175-3

Mikhael, J. G., and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS Comput. Biol.* 12:e1005062. doi: 10.1371/journal.pcbi.1005062

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Nambu, A. (2007). Globus pallidus internal segment. *Prog. Brain Res*. 160, 135–150. doi: 10.1016/S0079-6123(06)60008-3

Ott, T., and Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends Cogn. Sci.* 232, 13–234. doi: 10.1016/j.tics.2018.12.006

Ralph, D. E. (1968). Stimulus generalization among schizophrenics and normal subjects. *J. Abnorm. Psychol*. 73, 605–609. doi: 10.1037/h0026608

Reynolds, J. N., Hyland, B. I., and Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature* 413, 67–70. doi: 10.1038/35092560

Rummery, G., and Niranjan, M. (1994). *On-Line Q-Learning Using Connectionist Systems. Technical Report CUED/F-INFENG/TR 166*. Cambridge: Cambridge University Engineering Department

Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiol. Rev*. 95, 853–951. doi: 10.1152/physrev.00023.2014

Shepard, R. N., and Kannappan, S. (1991). "Connectionist implementation of a theory of generalization," in *Paper Presented at the Advances in Neural Information Processing Systems, Vol. 3.* Available Online at: http://papers.nips.cc/paper/351-connectionist-implementation-of-a-theory-of-generalization.pdf

Surmeier, D. J., Shen, W., Day, M., Gertler, T., Chan, S., Tian, X., et al. (2010). The role of dopamine in modulating the structure and function of striatal circuits. *Prog. Brain Res*. 183, 149–167. doi: 10.1016/S0079-6123(10)83008-0

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction. 2nd Edn*. Cambridge, MA: The MIT Press.

Tepper, J. M., Abercrombie, E. D., and Bolam, J. P. (2007). Basal ganglia macrocircuits. *Prog. Brain Res*. 160, 3–7. doi: 10.1016/S0079-6123(06)60001-0

Terashima, H., and Okada, M. (2012). "The topographic unsupervised learning of natural sounds in the auditory cortex," in *Paper Presented at the Advances in Neural Information Processing Systems, Vol. 25*. Available Online at: http://papers.nips.cc/paper/4703-the-topographic-unsupervised-learning-of-natural-sounds-in-the-auditory-cortex.pdf

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *Psychol. Rev. Monogr. Suppl.* 2, 1–8. doi: 10.1037/h0092987

van den Heuvel, M. P., Scholtens, L. H., de Reus, M. A., and Kahn, R. S. (2016). Associated microscale spine density and macroscale connectivity disruptions in schizophrenia. *Biol. Psychiatry* 80, 293–301. doi: 10.1016/j.biopsych.2015.10.005

Whittington, J. C. R., and Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250. doi: 10.1016/j.tics.2018.12.005

Wisniewski, M. G., Radell, M. L., Guillette, L. M., Sturdy, C. B., and Mercado, E. (2012). Predicting shifts in generalization gradients with perceptrons. *Learn. Behav*. 40, 128–144. doi: 10.3758/s13420-011-0050-6

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* 345, 1616–1620. doi: 10.1126/science.1255514

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.