EJN European Journal of Neuroscience

FENS Federation of European Neuroscience Societies

European Journal of Neuroscience, Vol. 42, pp. 2003–2021, 2015 doi:10.1111/ejn.12994

COMPUTATIONAL NEUROSCIENCE

Computing reward-prediction error: an integrated account of cortical timing and basal-ganglia pathways for appetitive and aversive learning

Kenji Morita[1] and Yasuo Kawaguchi[2,3,4]

[1]Physical and Health Education, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

[2]Division of Cerebral Circuitry, National Institute for Physiological Sciences, Okazaki, Japan

[3]Department of Physiological Sciences, SOKENDAI (The Graduate University for Advanced Studies), Okazaki, Japan

[4]Japan Science and Technology Agency, Core Research for Evolutional Science and Technology, Tokyo, Japan

Keywords: corticostriatal, direct pathway, dopamine, indirect pathway, reinforcement learning

Abstract

There are two prevailing notions regarding the involvement of the corticobasal ganglia system in value-based learning: (i) the direct and indirect pathways of the basal ganglia are crucial for appetitive and aversive learning, respectively, and (ii) the activity of midbrain dopamine neurons represents reward-prediction error. Although (ii) constitutes a critical assumption of (i), it remains elusive how (ii) holds given (i), with the basal-ganglia influence on the dopamine neurons. Here we present a computational neural-circuit model that potentially resolves this issue. Based on the latest analyses of the heterogeneous corticostriatal neurons and connections, our model posits that the direct and indirect pathways, respectively, represent the values of upcoming and previous actions, and up-regulate and down-regulate the dopamine neurons via the basal-ganglia output nuclei. This explains how the difference between the upcoming and previous values, which constitutes the core of reward-prediction error, is calculated. Simultaneously, it predicts that blockade of the direct/indirect pathway causes a negative/positive shift of reward-prediction error and thereby impairs learning from positive/negative error, i.e. appetitive/aversive learning. Through simulation of reward-reversal learning and punishment-avoidance learning, we show that our model could indeed account for the experimentally observed features that are suggested to support notion (i) and could also provide predictions on neural activity. We also present a behavioral prediction of our model, through simulation of inter-temporal choice, on how the balance between the two pathways relates to the subject's time preference. These results indicate that our model, incorporating the heterogeneity of the cortical influence on the basal ganglia, is expected to provide a closed-circuit mechanistic understanding of appetitive/aversive learning.

Introduction

The corticobasal ganglia system has been shown to be centrally involved in value-based learning (Robbins & Everitt, 1996; Hikosaka *et al.*, 2006; Graybiel, 2008). A prevailing notion is that the two main pathways of the basal ganglia (BG), called the 'direct' and 'indirect' pathways (Gerfen & Surmeier, 2011), are specialized for learning from good (appetitive) and bad (aversive) outcomes to reinforce 'Go' and 'No-Go' responses, respectively (Frank *et al.*, 2004; Frank, 2005; Hong & Hikosaka, 2011; Collins & Frank, 2014; Nakanishi *et al.*, 2014). This notion stands on another popular notion that dopamine represents the reward-prediction error (RPE; Montague *et al.*, 1996; Schultz *et al.*, 1997), assuming that positive/negative RPE caused by appetitive/aversive outcomes is encoded by dopamine increase/decrease, which differentially strengthens the direct/indirect pathway through plasticity. On the other hand, neural circuit mechanisms enabling the midbrain dopamine neurons to calculate RPE are usually not considered to be related to the BG-pathway-specialization hypothesis.

However, the dopamine neurons in fact receive major inputs from the BG. In addition to the direct inputs from striatal neurons in the striosomes/patches (Gerfen, 1984; Fujiyama *et al.*, 2011; Watabe-Uchida *et al.*, 2012), inputs through the BG direct and indirect pathways would also have a significant contribution, given that neurons in the substantia nigra pars reticulata (SNr), an output nucleus of the BG, have been suggested to send influential inputs to the dopamine neurons via axon collaterals (Tepper *et al.*, 1995; Tepper & Lee, 2007). Indeed, although there are models of RPE calculation that do not assume the involvement of either BG pathway (e.g. Hazy *et al.*, 2010), others incorporate either both (e.g. Doya, 2002; Aggarwal *et al.*, 2012) or at least one of the BG pathways. Therefore, an emerging question is whether any of the latter models are in line with, or could even account for, experimental observations that have been argued to be evidence for the specialization of the BG pathways for appetitive and aversive learning. However, this question has been rarely addressed.

Correspondence: Kenji Morita, as above.
E-mail: morita@p.u-tokyo.ac.jp

Received 21 April 2015, revised 11 June 2015, accepted 17 June 2015

Recently, we have proposed a model of the mechanism of RPE calculation (Morita *et al.*, 2012, 2013; Morita, 2014). Whereas most previous models disregarded the heterogeneity of cortical cells, our model incorporates recently revealed features of two types of corticostriatal neurons (Morishima & Kawaguchi, 2006; Morishima *et al.*, 2011), as well as their differential activation of the two BG pathways that have been suggested by anatomical results (Lei *et al.*, 2004; Reiner *et al.*, 2010; Deng *et al.*, 2015) and also by the latest computational analyses (Morita, 2014) of physiological results on short-term plasticity. Because our model incorporates the two BG pathways, the abovementioned question arises of whether our model accords with, or could even account for, the observations suggested to support the appetitive/aversive specialization of the pathways. Here we address this issue, targeting two studies (Hikida *et al.*, 2010; Yawata *et al.*, 2012) and presenting predictions on neural activity. We also provide a behavioral prediction, through simulating inter-temporal choice, about how the strengths of the two pathways could affect the subject's time preference.

Materials and methods

Corticostriatal temporal difference model

We describe the architecture of our model, named the corticostriatal temporal difference (CS-TD) model (Fig. 1), together with the experimental findings and analyses on which the model is based.

Corticostriatal neurons

In the neocortex including the frontal areas, there exist two types of corticostriatal neurons, called crossed-corticostriatal (CCS) cells and corticopontine/pyramidal-tract (CPn/PT) cells (Cowan & Wilson,

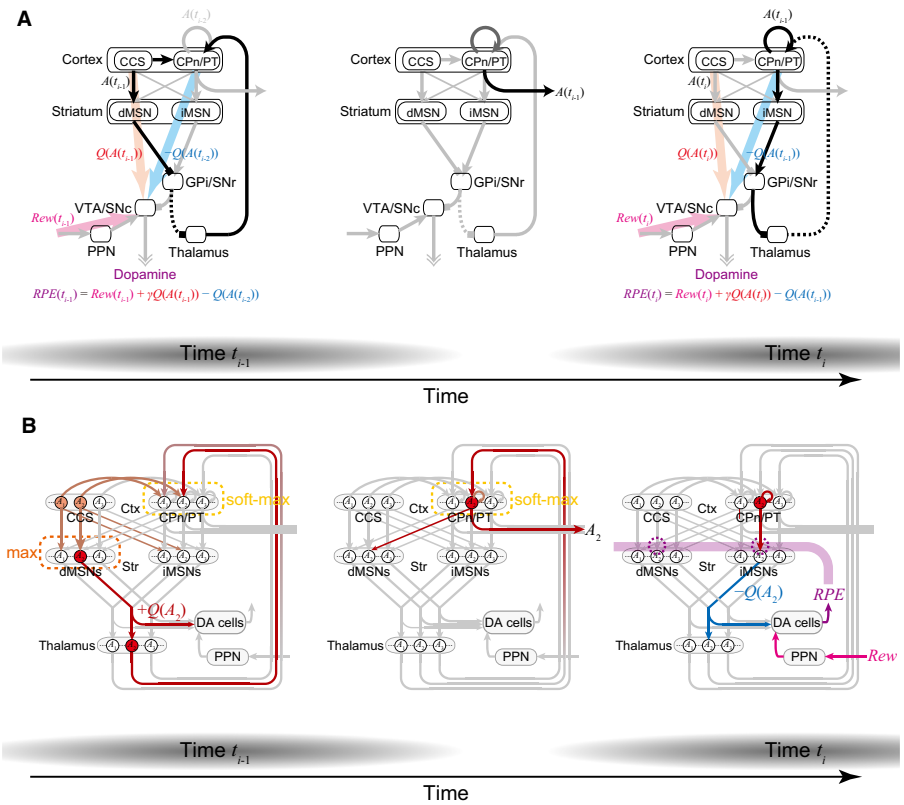

FIG. 1. The CS-TD model for the mechanism of reward-prediction error (RPE) calculation and RPE-based reinforcement learning. The panels in this figure were taken from the original figure in Morita (2014) with modifications. (A) Two types of corticostriatal cells [crossed-corticostriatal (CCS) cells and corticopontine/pyramidal-tract (CPn/PT) cells] represent the current/upcoming action (action candidate) and previous/executed action [e.g. $A(t_i)$ and $A(t_{i-1})$ at time t_i (right panel)], respectively, by virtue of CCS→CPn/PT unidirectional projections and strong recurrent excitation among CPn/PT cells. CCS and CPn/PT cells predominantly activate the direct-pathway medium spiny neurons (dMSNs) and the indirect-pathway medium spiny neurons (iMSNs), respectively, and thus dMSNs/iMSNs represent the value of current/previous actions [$Q(A(t_i))/Q(A(t_{i-1}))$]. The dopamine (DA) neurons in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) are up-regulated/down-regulated by dMSNs/iMSNs via the output nuclei of the BG (internal segment of globus pallidus (GPi) / substantia nigra pars reticulata (SNr)), and also receive information about obtained reward [$Rew(t_i)$] from the brainstem pedunculopontine nucleus (PPN); thereby $RPE(t_i) = Rew(t_i) + Q(A(t_i)) - Q(A(t_{i-1}))$ is calculated. Notably, RPE is calculated at every time step (left: t_{i-1}; right: t_i). The black/gray contrast of the lines indicates a selection process over time (black, currently ON; gray, currently OFF). The fuzzy gray areas surrounding the time steps (t_{i-1} and t_i) at the bottom indicate how time steps can be defined and how well the description using the discrete time steps can approximate the actual continuous-time dynamics remains to be examined. (B) Higher-resolution diagram of the model, illustrating the mechanism of action selection, execution, and update of action values. Ctx, Str, and Rew represent the cortex, the striatum, and reward respectively. In each neural population (CCS, CPn/PT, dMSNs, iMSNs, and thalamus), subpopulations of neurons corresponding to different actions are illustrated by small circles. If there are multiple action candidates (in the case shown in the figure, A_1 and A_2), competition occurs in the CCS→dMSNs circuit and a dMSN subpopulation corresponding to the maximum-valued action (in the figure, A_2) eventually wins and represents $\max_j\{Q(A_j)\}$. CPn/PT subpopulations receive information about the values of the corresponding action candidates represented, at least transiently, by dMSN subpopulations via the direct pathway and thalamus, and compete with each other. One subpopulation eventually wins, with the probability depending on the value of each action (i.e. soft-max action selection), and the corresponding action (in the figure, A_2) is executed and its value is updated according to RPE.

1994; Reiner *et al.*, 2010; Hirai *et al.*, 2012; Shepherd, 2013). It has been shown that there are unidirectional connections from CCS cells to CPn/PT cells (Morishima & Kawaguchi, 2006) and strong, facilitatory recurrent excitation among CPn/PT cells with a relatively high degree of reciprocal connections (Morishima *et al.*, 2011). From these properties, it is indicated that information can be sent from CCS cells to CPn/PT cells, and kept there as self-sustained activity held by recurrent excitation. Therefore, we assumed in the model that the subject's current/upcoming action (or action candidate) [$A(t_i)$] and previous/executed action [$A(t_{i-1})$] are represented by single subpopulations of CCS and CPn/PT cells, respectively.

Corticostriatal connections

Regarding the connectivity between the two types of corticostriatal cells and the two types of striatal projection neurons (medium spiny neurons) projecting to the direct and indirect pathways (dMSNs and iMSNs), anatomical studies have shown that there exist CCS→dMSN and CPn/PT→iMSN connection preferences (Lei *et al.*, 2004; Reiner *et al.*, 2010). The activation preferences expected from these anatomical results were, however, not observed in experiments with brief electrical or optical stimulation of cortical neurons/axons (Ballion *et al.*, 2008; Kress *et al.*, 2013). Nonetheless, recent model-fitting analyses (Morita, 2014) of experimental results on short-term synaptic plasticity (Kreitzer & Malenka, 2007; Ding *et al.*, 2008) have suggested that repetitive spikes of CCS cells and CPn/PT cells predominantly activate dMSNs and iMSNs, respectively, by virtue of a combination of the anatomical connection preferences and presumably synapse type-dependent short-term plasticity. Also, a very recent anatomical study (Deng *et al.*, 2015) has re-examined the corticostriatal connections. As a result, it was confirmed that there exist CCS→dMSN and CPn/PT→iMSN connection preferences for synapses on dendritic spines. At the same time, however, it was also found, for the first time, that no such biases, but rather a CPn/PT→dMSN preference, exist for synapses on dendritic shafts. This new finding can reconcile the abovementioned discrepancy between the anatomical and physiological results. Moreover, given that spines are the crucial sites for dopamine-dependent plasticity (Yagishita *et al.*, 2014), whereas inputs to shafts could potentially represent broader contextual information, the CCS→dMSN and CPn/PT→iMSN preferences for synapses on spines can be very relevant for learning of the values of specific actions. Based on these considerations, we assumed in the model that dMSNs and iMSNs signal the values (reward expectations) of the current and previous actions [$A(t_i)$ and $A(t_{i-1})$], which are assumed to be represented by CCS and CPn/PT cells, respectively, as mentioned above (see below for more details).

Activity of striatal medium spiny neurons

$$x_{\mathrm{dMSN}}(t_i) = f_d(Q(A(t_i))) \text{ (unless there are two possible actions at } t_i) \quad (1)$$

$$x_{\mathrm{iMSN}}(t_i) = f_i(Q(A(t_{i-1}))) \quad (2)$$

$x_{\mathrm{dMSN}}(t_i)$ and $x_{\mathrm{iMSN}}(t_i)$ represent the population activity (= activity of an active subpopulation) of dMSNs and iMSNs at time t_i, respectively. $A(t_i)$ and $A(t_{i-1})$ are an action (or action candidate) being taken at time t_i and an action executed at time t_{i-1}, respectively (the case with multiple action candidates is described below). $Q(A)$ represents the predicted value of action A, and is assumed to be represented by the strength of the input to medium spiny neurons (MSNs), reflecting the strength of synaptic connections between A-corresponding subpopulations of CCS cells and dMSNs and that of CPn/PT cells and iMSNs. This assumption was made based on experimental findings that action values and chosen values are coded in the striatal neurons (Samejima *et al.*, 2005; Lau & Glimcher, 2008; Ito & Doya, 2009; Kim *et al.*, 2009; Seo *et al.*, 2012), and is bolstered by a recently developed formal framework for action discovery (Gurney *et al.*, 2013). $Q(A)$ was initially set to 0 for arbitrary action A in most cases, except for one of the two types of simulations of punishment-avoidance task (see below), and was updated according to a rule described below. f_d and f_i are functions representing the transformation from the strength of synaptic inputs (taking into account the connection strength that can change through synaptic plasticity as described below) to the output activity, and are assumed to be the threshold-linear (rectifying) function with the threshold and the slope set to 0 and 1, respectively (i.e. $f_d(z) = f_i(z) = 0$ (if $z \leq 0$) or z (if $z > 0$)) in the case without blockade of BG pathway. Notably, we hypothesized that different dMSN or iMSN subpopulations correspond to different actions, but the variable x_{dMSN} or x_{iMSN} is assumed to represent the activity of the active dMSN or iMSN subpopulation, respectively. In the case where there exist two action candidates at t_i, in particular, A_1 and A_2, we assumed the following:

$$x_{\mathrm{dMSN}}(t_i) = \max\{f_d(Q(A_1)), f_d(Q(A_2))\} \quad (3)$$

This max (i.e. maximum) operation is assumed to be realized through competitive neural dynamics at the CCS→dMSNs circuit, possibly via effective lateral inhibition on the dendrites of MSNs (Moyer *et al.*, 2014) and feed-forward inhibition through fast-spiking interneurons (Plenz & Kitai, 1998; Mallet *et al.*, 2005; Gittis *et al.*, 2010).

Activity of pedunculopontine nucleus neurons

$$x_{\mathrm{PPN}}(t_i) = r \text{ (at the time of reward) or } 0 \text{ (otherwise)} \quad (4)$$

$x_{\mathrm{PPN}}(t_i)$ represents the activity of a population of pedunculopontine nucleus (PPN) neurons that represent the obtained reward (cf. Okada *et al.*, 2009). r represents the size (amount) of reward.

Response of dopamine neurons

We assumed that the dopamine neurons receive net positive and negative influences from dMSNs and iMSNs, respectively, via the output nuclei of BG, and also receive inputs from the PPN neurons representing obtained reward. Specifically, we assumed the following:

$$x_{\mathrm{DA}}(t_i) = x_{\mathrm{PPN}}(t_i) + \gamma x_{\mathrm{dMSN}}(t_i) - x_{\mathrm{iMSN}}(t_i) \quad (5)$$

where $x_{\mathrm{DA}}(t_i)$ represents the response of dopamine neurons (or resulting dopamine release) compared with its baseline level [$x_{\mathrm{DA}}(t_i)$ can be negative, in response to aversive events or cues predicting them; cf. Schultz *et al.* (1997), Ungless *et al.* (2004), Hart *et al.* (2014)], and γ is a parameter representing the relative strength of the direct pathway over the indirect pathway, which corresponds to the time discount factor according to the CS-TD model. γ was set to 0.75 in the present work. The terms '$+ \gamma x_{\mathrm{dMSN}}(t_i)$' and '$- x_{\mathrm{iMSN}}(t_i)$' were assumed to come through the dMSNs→internal segment of the globus pallidus (GPi)/SNr→ventral tegmental area/substantia nigra pars compacta (dis-inhibition, net positive) and the iMSNs→external segment of the globus pallidus (GPe)→[subthalamic nucleus (STN)]→GPi/SNr→ventral tegmental area (VTA)/substantia nigra

pars compacta (SNc) (dis-dis-inhibition, net negative), respectively. As such, the terms '+ $\gamma x_{\mathrm{dMSN}}(t_i)$' and '− $x_{\mathrm{iMSN}}(t_i)$' are assumed to be realized by changes in the amount of inhibition that the dopamine neurons receive from the GPi/SNr. Related to this, it is noteworthy that a decrease in inhibition (i.e. dis-inhibition) has been shown to be able to cause burst firing of the dopamine neurons (Lobb *et al.*, 2011). The involvement of these multi-synaptic routes from MSNs through the direct and indirect pathways to the dopamine neurons in the RPE calculation has also been suggested by other researchers (e.g. Doya, 2002; Aggarwal *et al.*, 2012) although the existence of the multi-synaptic routes remains to be empirically demonstrated. Notably, we assumed that there exist subpopulations of GPi/SNr neurons, each of which corresponds to a single action, so that the entire GPi/SNr represents action as a vector, whereas VTA/SNc neurons uniformly represent RPE as a scalar. Because we assumed the temporal difference between the direct and indirect pathways, even if a single GPi/SNr neuron receives inputs from both dMSNs and iMSNs (corresponding to the same action), the timings of these inputs are expected to differ (the former precedes the latter). In fact, it has been suggested (Sato & Hikosaka, 2002) that there may exist two subpopulations of SNr neurons that are downstream of the direct and indirect pathways. It should also be noted that the direct inputs from the striosomal neurons to the dopamine neurons were not assumed to be included in the above description of the model; however, such inputs could also contribute to the term '− $x_{\mathrm{iMSN}}(t_i)$' if the striosomal neurons receive preferential inputs from CPn/PT cells as has been discussed (but not yet resolved; Crittenden & Graybiel, 2011; Shepherd, 2013), as discussed previously (Morita *et al.*, 2012). A further issue is the possible difference (Smith *et al.*, 2013) in the circuit architecture, including whether the direct and indirect pathways are separated, between the dorsal striatum and the ventral striatum/nucleus accumbens (NAc), which was the target of the reversal learning experiment (Yawata *et al.*, 2012) that we modeled. This point needs to be clarified in the future (Smith *et al.*, 2013; Nakanishi *et al.*, 2014).

Probabilistic action selection

If there are multiple action candidates (A_1 and A_2), the CPn/PT subpopulations are assumed to receive information about the values of the corresponding action candidates represented, at least transiently, by dMSN subpopulations via the direct pathway and the thalamus, and compete with each other {it is assumed that the A_1-corresponding and A_2-corresponding dMSN subpopulations are both transiently active, reflecting $Q(A_1)$ and $Q(A_2)$, but only a subpopulation corresponding to action with the larger value survives after the transient period and affects the calculation of RPE [see Morita *et al.* (2013) for more explanation for this assumption]}. It is then assumed that one CPn/PT subpopulation eventually wins, with the probability depending on the value of each action [i.e. soft-max action selection; cf. Wang (2002), Soltani & Wang (2008), Hunt *et al.* (2012), Jocham *et al.* (2012)], and the corresponding action is selected and executed. Specifically, we assumed that the following equations represent the soft-max action selection:

$$\mathrm{Prob}(A_1) = 1/(1+\exp(-(f_d(Q(A_1)) - f_d(Q(A_2)))/\varepsilon)), \text{and} \quad (6)$$

$$\mathrm{Prob}(A_2) = 1/(1+\exp(-(f_d(Q(A_2)) - f_d(Q(A_1)))/\varepsilon)) = 1 - \mathrm{Prob}(A_1) \quad (7)$$

where $\mathrm{Prob}(A_1)$ and $\mathrm{Prob}(A_2)$ represent the probability of choosing action A_1 and A_2, respectively (bottom inset in Fig. 2C). ε is a parameter determining the 'flatness' of the sigmoidal (soft-max) function, and it represents the degree of exploration (over exploitation) upon action selection. ε was set to 1/8 (Figs 2C inset and 7F) in the simulations shown in Figs 3–6, 7C–G, and 9A, and set to 2/3 (Fig. 8D) in the simulations shown in Figs 8 and 9B. Notably, by assuming the coupling of 'max' and 'soft-max' operations in dMSNs and CPn/PT cells, respectively, our model can approximately implement a reinforcement learning algorithm called Q-learning (Watkins, 1989), which has been suggested to be implemented in the cortico-BG-midbrain system (Roesch *et al.*, 2007; for more details, see Morita *et al.*, 2013; Morita,

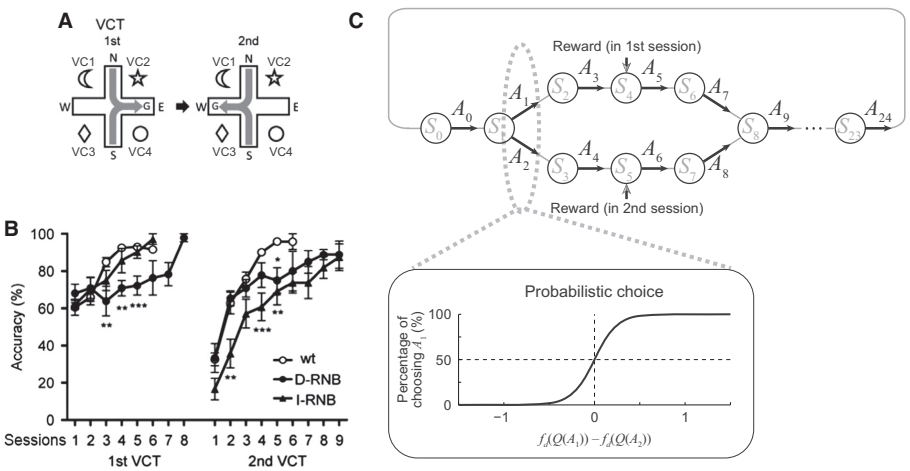

FIG. 2. Distinct leaning impairments caused by blockade of the direct or indirect pathway (d-block and i-block) of the BG, more specifically, of the NAc, in a reversal learning task. Experimental results and schematic illustration of our simulation of a simplified task. The panels in A and B were taken from Yawata *et al.* (2012) with permission. (A) Cue-target/action association task with a contingency reversal used in Yawata *et al.* (2012). Reversal occurred between 1st VCT (visual cue task) and 2nd VCT. VC1–VC4, visual cue 1–visual cue 4; G, goal. (B) Experimental results for the task shown in A (Yawata *et al.*, 2012). The graph shows the percentage of choosing action that leads to reward (vertical axis) along with learning sessions (horizontal axis; each session contains 12 trials) in the control [wild-type (wt)], d-block [D-RNB (direct-pathway reversible neurotransmission blocking)], and i-block (indirect-pathway reversible neurotransmission blocking, I-RNB) mice. The error bars represent the mean ± SEM [n = 11 (wt), n = 6 (D-RNB), and n = 7 (I-RNB)]. (C) Simulated simplified reversal learning task, in which state/action-reward contingency is reversed between the first and second session. The subject's behavior is represented by sequential state transitions, and calculation of RPE according to the CS-TD model (as shown in Fig. 1) is assumed to be executed at each state, with the effect of d-block or i-block incorporated. Bottom inset: probabilistic action selection depending on the predicted values of A_1 and A_2 [$Q(A_1)$ and $Q(A_2)$] assumed in the model. f_d is the assumed neuronal input–output transformation function of dMSNs. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, respectively.

2014). It is also notable that these mechanisms assumed in our model could be said to be largely in line with the arguments that the cortico-BG system implements action selection (Mink, 1996; Redgrave *et al.*, 1999; Hikosaka *et al.*, 2000; Humphries *et al.*, 2006), whereas specific roles of different subpopulations of cortical and striatal neurons are newly assumed. Competitive processes in the wider BG circuit, which are not included in our model, may also contribute to action selection, and examining this point in relation to our model is an important future issue.

Dopamine-dependent plastic changes of the crossed-corticostriatal→direct-pathway medium spiny neuron and corticopontine/pyramidal tract→indirect-pathway medium spiny neuron transmissions

We assumed that the RPE-representing dopamine signal causes plastic modification of the strength/efficacy of the CCS→dMSN and CPn/PT→iMSN synaptic transmissions so that the value of the executed action stored in these connections is updated according to RPE. Specifically, we assumed the following:

$$Q(A(t_{i-1})) \rightarrow Q(A(t_{i-1})) + \alpha x_{\mathrm{DA}}(t_i) \quad (8)$$

where α represents the learning rate ($0 < \alpha < 1$). Notably, the CS-TD model assumes that the two (CCS→dMSN and CPn/PT→iMSN) corticostriatal transmissions are plastically modified in the same direction (rather than in opposite directions); see the Discussion for the potential validity of this assumption. The learning rate α was set to 0.05 unless otherwise described. We also examined cases with reward history-dependent adaptive modulation of the learning rate, in which we assumed that the learning rate is initially 0.05, and increases to 0.2 whenever the subject experiences four consecutive rewarded trials followed by two consecutive unrewarded trials, and then decays exponentially with the time constant of 20 trials:

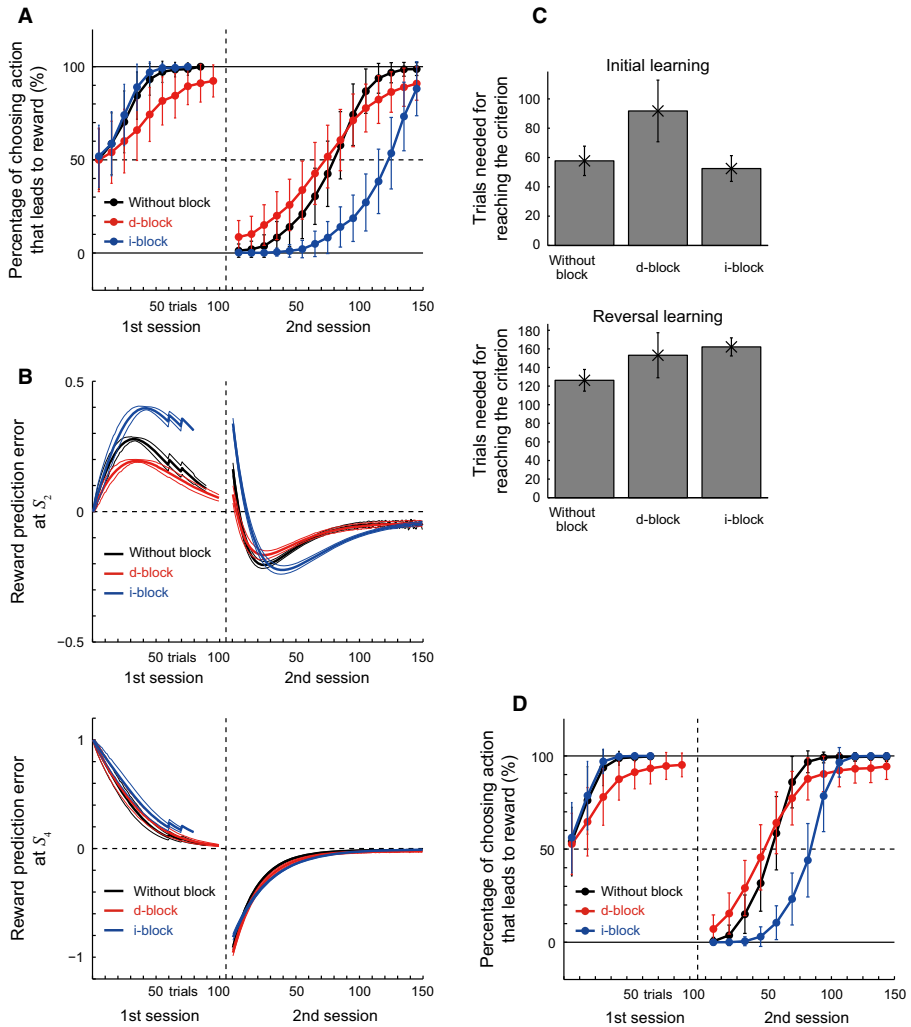

FIG. 3. Results of the simulations of the reversal learning task by the CS-TD model. (A) The percentage of choosing action that leads to reward (vertical axis) along with trials (horizontal axis), calculated with the bin size = 10 trials. The error bars represent the mean ± SD across simulations. The black, red, and blue colors indicate the control (without blockade), d-block, and i-block cases, respectively (the same color policy is applied to the subsequent figures). (B) RPE at state S_2 (in Fig. 2C) (top panel) or S_4 (bottom panel). The thick and thin lines represent the mean ± SD across simulations in which S_2 and S_4 were visited (i.e. A_1 was chosen). (C) Number of trials needed to reach the performance criterion (≥ 95% rewarded choice in the last 20 trials) before (top panel) or after (bottom panel) the contingency reversal in the control (left bars), d-block (middle bars), and i-block (right bars) cases. Whether the criterion was reached or not was evaluated at every 10 trials in each session, and the error bars represent the mean ± SD across simulations. (D) Results of the simulations with the learning rate doubled. The configurations are the same as those in A.

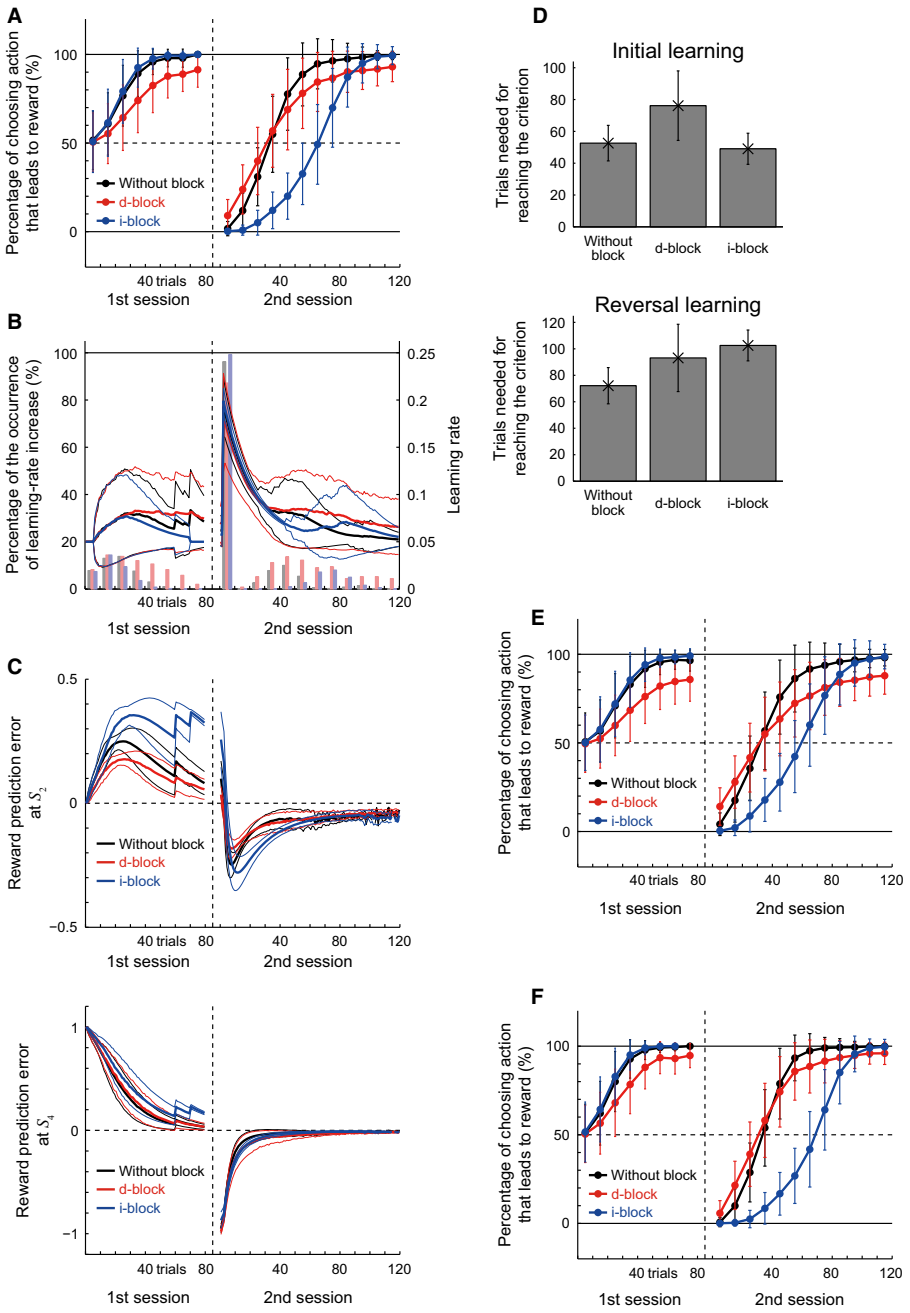

FIG. 4. Results of the simulations of the elaborated model incorporating reward history-dependent adaptive modulation of the learning rate, in which the learning rate was assumed to increase whenever the subject experienced four consecutive rewarded trials followed by two consecutive unrewarded trials (modeling a change in the reward environment) and then exponentially decay to the baseline value. The configurations (except for B) are the same as those in Fig. 3. (A) The percentage of choosing action that leads to reward. (B) Changes in the learning rate along with trials (horizontal axis) during the reversal learning task. The histograms represent the frequency of occurrence of the increase of the learning rate across simulations. The thick lines and the thin lines above and below the thick lines with the same color represent the mean ± SD of the learning rate across simulations. The black, red, and blue bars/lines indicate the control, d-block, and i-block cases, respectively. (C) RPE at state S_2 (in Fig. 2C) (top panel) or S_4 (bottom panel). (D) Number of trials needed to reach the performance criterion before (top panel) or after (bottom panel) the contingency reversal. Simulation results for the percentage of choosing action that leads to reward in the model with a larger (E) or smaller (F) degree of exploration upon action selection.

$$\alpha = 0.05 + 0.15\ \exp(-(N_{\mathrm{ap}} - 1)/20) \qquad (9)$$

where N_{ap} represents the number of trials after the last experience of the sequence of four-rewarded–two-unrewarded trials (N_{ap} = 1 for the trial just following the sequence). For the simulations of the punishment-avoidance task, we also examined cases assuming that punishment induces an increase of the learning rate (see below). Notably, dopamine has been suggested to modulate both the plasticity of corticostriatal synapses and responsiveness of striatal neurons (Gerfen & Surmeier, 2011), presumably through (mainly) phasic and tonic actions, respectively, but only the former was incorporated into the model described here.

Simulation of behavioral tasks with pathway blockade

By using the CS-TD model, we simulated the operation of the corticobasal ganglia circuit and the resulting behavior in three tasks: (i) a

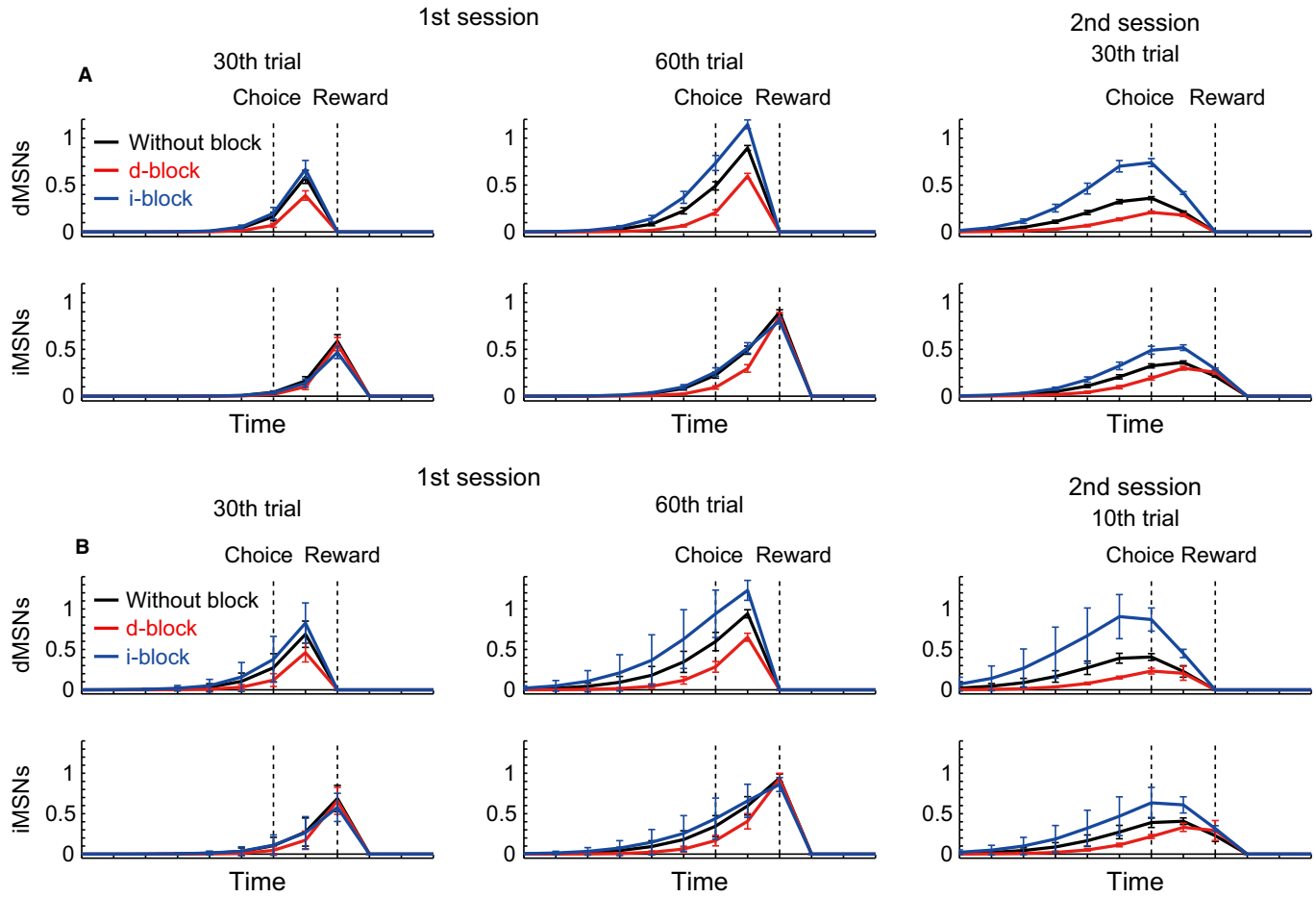

FIG. 5. Simulated population activity of MSNs in the reversal learning task. Output activity taking into account the effects of the pathway blockade is shown. (A) Results of the simulations with the original model used for Fig. 3A–C. Population activity of dMSNs (top) and iMSNs (bottom) in the 30th (left) and 60th (middle) trials of the first session and the 30th trial of the second session (right) in which action A_1 (in Fig. 2C) was chosen. The error bars represent the mean ± SD across simulations in which A_1 was chosen, out of the total 500 simulations. The black, red, and blue colors indicate the control (without blockade), d-block, and i-block cases, respectively. The two vertical dashed lines labeled with 'Choice' and 'Reward' indicate the timings corresponding to S_1 (taking A_1) and S_4 (receiving reward and taking A_5) in Fig. 2C, respectively. (B) Results of the simulations with the elaborated model incorporating reward history-dependent adaptive modulation of the learning rate used for Fig. 4A–D. The right panels show the population activity in the 10th (rather than the 30th) trial of the second session.

cue-target/action association task with a contingency reversal (referred to as 'reversal learning task'), which is a simplification of a task used in Yawata *et al.* (2012) (see the Results for details about the simplification); (ii) a punishment-avoidance task, which was intended to model avoidance of electric footshock (Hikida *et al.*, 2010); and (iii) an inter-temporal choice task, which is a virtual task that we consider in order to present predictions of the CS-TD model. For all of the tasks, we assumed that a diagram of action-dependent state transitions that defines the task (or the 'model' of the environment; Fig. 2C for the reversal learning task, Fig. 6A and E for the punishment-avoidance task, and Fig. 7B for the inter-temporal choice task) is represented in the subject's cerebral cortex, in particular in the orbitofrontal cortex (Wilson *et al.*, 2014) that projects to the NAc. At each presumed 'time-step' (t_i), the subject is assumed to exist at one of the states (circles in Figs 2C, 6A or E, or 7B) and is taking an action $A(t_i)$; if there are two action candidates (A_1 or A_2), the subject is choosing one of them depending on their values according to the soft-max rule described above (Fig. 2C, bottom inset). We assumed that, at each time-step/state, calculation of the RPE and RPE-dependent update of the value of previous action, together with action selection and execution, is conducted in the cortico basal ganglia circuit according to the CS-TD model. Notably, although we assumed that the 'model' of the environment is represented in the orbitofrontal cortex as mentioned above, we did not assume that the existence of the association-reversal itself was included in the 'model' of the reversal learning task (a potential rationale of this is that presumably the existence of the reversal could be incorporated into the 'model' only after the subject experiences the reversal many times). Also, we assumed that the cortico-BG circuit considered in the CS-TD model implements the so-called 'model-free' decision making (action selection) based on cached values acquired through RPE-based learning, rather than the 'model-based' decision making based on real-time calculation of the expected total future values of action candidates through mental simulations of state transitions (Daw *et al.*, 2005).

The reversible neurotransmission blocking used in Yawata *et al.* (2012) and Hikida *et al.* (2010) is a method to express a bacterial toxin, which abolishes neurotransmitter release from synaptic vesicles, specifically in dMSNs or iMSNs by using the transgenic technique, and therefore we modeled the reversible neurotransmission blocking by reducing the output of (rather than the input to) MSNs. Specifically, we simulated the conditions with blockade of either the

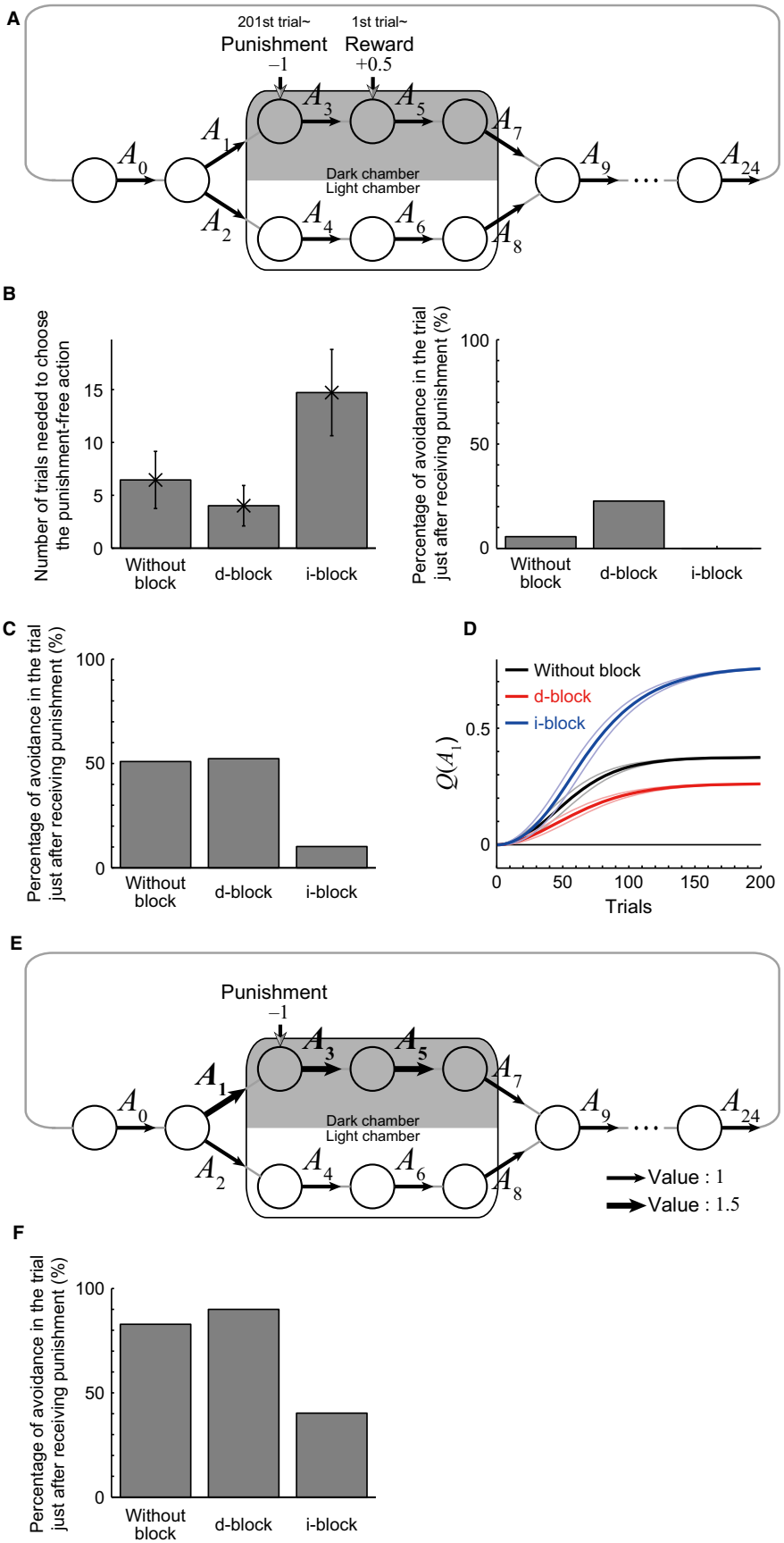

FIG. 6. Potential explanations of the experimentally suggested critical involvement of the indirect pathway in punishment-avoidance learning by the CS-TD model. (A) Simulated punishment-avoidance learning task, in which the subject receives reward (+0.5) after taking action A_3 (corresponding to the mouse's preference for a dark chamber), but, after 200 trials, the subject receives punishment (−1) after taking action A_1 (corresponding to the electric footshock). (B) Results of simulations with the same parameters used in Fig. 3A–C (the learning rate was constant at 0.05). Left panel: the number of trials needed to choose the punishment-free action (A_2) for the first time after receiving the punishment in the control (without blockade) (left), d-block (middle), and i-block (right) cases. The error bars represent the mean ± SD across simulations. Right panel: the across-simulation percentage of avoidance (choosing A_2) in the trial just after the subject received punishment for the first time (i.e. the 202nd trial). (C) The across-simulation percentage of avoidance in the 202nd trial in a different set of simulations assuming that punishment induces an increase of the learning rate (0.05–0.5). (D) Time evolution of the value of action A_1 during the appetitive learning phase (i.e. 1st–200th trials) in the control (without blockade) (black), d-block (red), and i-block (blue) cases. The error bars represent the mean ± SD across simulations. (E) A different way of simulation of the avoidance of punishment, where appetitive reward learning was not assumed but instead action values were set *a priori*, with the actions corresponding to the dark chamber (i.e. A_1, A_3, and A_5) having larger values (1.5) than the others (1) (representing the preference for the darkness). (F) The across-simulation percentage of avoidance (choosing A_2) in the trial just after receiving punishment in the simulations of the setting shown in E.

direct or indirect pathway (referred to as d-block or i-block) of the NAc in Yawata *et al.* (2012) or of the striatum including the NAc in Hikida *et al.* (2010) by reducing the slope of the input–output transformation function f_d or f_i, respectively (for the region with positive input) from the original value of 1–0.7 [i.e. $f_d(z)$ or $f_i(z) = 0$ (if $z \leq 0$) or $0.7z$ (if $z > 0$)]. The reduction of the slope to 0.7 rather than to 0 was meant to represent that the blockade in the experiments was presumably restricted to parts of the striatum [in particular the NAc in Yawata *et al.* (2012)] and the remaining parts of the striatum were presumably spared (it was also assumed that representations of actions are distributed across striatal regions). We have confirmed that moderately changing the degree of reduction of the slope (0.8 or 0.6) did not greatly change the qualitative tendency of the results shown in Figs 3A, 4A, 6B left, C and F, 7C, 8A and 9A top and B top, although in general the tendency was weaker or stronger when the reduced slope (f_d or f_i) was 0.8 or 0.6, respectively, than when the slope was 0.7 and, in particular, the impairment of avoidance by i-block was considerably weakened when the slope was 0.8 (data not shown). More severe reduction of f_d or f_i can change the tendency (reaching the performance criterion for the reversal learning task described below can be unachieved); the assumed partial (rather than complete) reduction of f_d or f_i can correspond to potentially occurring functional compensation by other parts of the striatum/BG and/or other brain regions.

For the reversal learning task (Fig. 2C), we assumed that reward (size 1) is obtained at state S_4 if the subject chooses A_1 in the first part of the task ['reward (in 1st session)' in Fig. 2C]. After the performance reached a certain criterion, specifically the percentage of choosing A_1 in the last 20 trials became ≥ 95%, the action-reward contingency was assumed to be reversed so that reward is obtained at state S_5 if the subject chooses A_2 ['reward (in 2nd session)' in Fig. 2C]. Whether the criterion was reached or not was evaluated at every 10 trials in the first session and also in the second session, and at least 60 trials were performed in the first session, regardless of the performance. For the punishment-avoidance task, we conducted two different methods of simulations (Figs 6A–F). For the first method (Fig. 6A–D), we considered a similar setting to the simulations of the reversal learning task, but introducing punishment, modeled by negative reward (−1), after 200 trials of initial appetitive reward (+0.5) learning (which was assumed so that the preference for the dark chamber appears) instead of the contingency reversal of positive reward. The learning rate was assumed to be either constant at 0.05 (Fig. 6B) or initially at 0.05 and increased to 0.5 upon receiving punishment (Fig. 6C). For the second method (Fig. 6E and F), appetitive reward learning was not assumed but instead action values were set *a priori*, with the actions corresponding to the dark chamber having larger values (1.5) than the other actions (1). The learning rate was assumed to be initially 0 and increased to 0.5 upon receiving punishment. For the inter-temporal choice task (Fig. 7B), the size of the immediate reward was always set to 1, and the size of the delayed reward was varied for 1, 2,. . .8. By using MATLAB (MathWorks Inc., Natick, MA, USA), we conducted 500 simulations for each condition, where different simulation runs are expected to differ in the sequence of selected actions across trials because of the assumed stochasticity at the softmax action selection in the model.

Results

The corticostriatal temporal difference model explains the apparent Go/No-Go specialization of the direct/indirect pathway

We asked whether our CS-TD model can account for experimental results that suggest specialization of the direct and indirect pathways for appetitive (Go) and aversive (No-Go) learning. As representatives of such results, we considered two studies. The first one is the study showing that selective blockade of the direct or indirect pathway causes distinct learning impairment in reward-reversal learning (Yawata *et al.*, 2012). Specifically, the authors selectively blocked transmission of either the direct or indirect pathway of the NAc (referred to as 'd-block' or 'i-block', respectively), and examined how the blockade affects learning in the cue-target/action association task with a contingency reversal or a rule change. We consider the former (the task with a contingency reversal; Fig. 2A), which consists of a pair of goal-choosing maze tasks (subtasks): (i) given visual cues VC1 and VC2 in front (starting from S), choose (i.e. turn to) the direction of VC2 to get reward, and (ii) given VC3 and VC4 in front (starting from N), choose (turn to) the direction of VC4 to get reward; after about 72–96 trials (12 trials/session × ~6–8 sessions), both of these cue–reward contingencies are reversed (Fig. 2A, right). These two subtasks were imposed on the subjects (mice) in pseudo-random order, and their behavioral data [fraction of rewarded response (accuracy)] were pooled. As shown in Fig. 2B, the authors (Yawata *et al.*, 2012) revealed that d-block and i-block selectively impair the initial learning and the early phase of reversal learning, respectively. These two learning processes can basically be regarded as appetitive and aversive learning, respectively, and the authors of the study have argued that their result is explicit experimental evidence for the specialization of the direct/indirect pathways for Go/No-Go learning.

We conjectured that our CS-TD model could account for these results without assuming specialization of the pathways for learning types. According to the CS-TD model (Fig. 1), dMSNs and iMSNs presumably up-regulate and down-regulate the dopamine neurons, respectively, and thus d-block or i-block is expected to cause a neg-

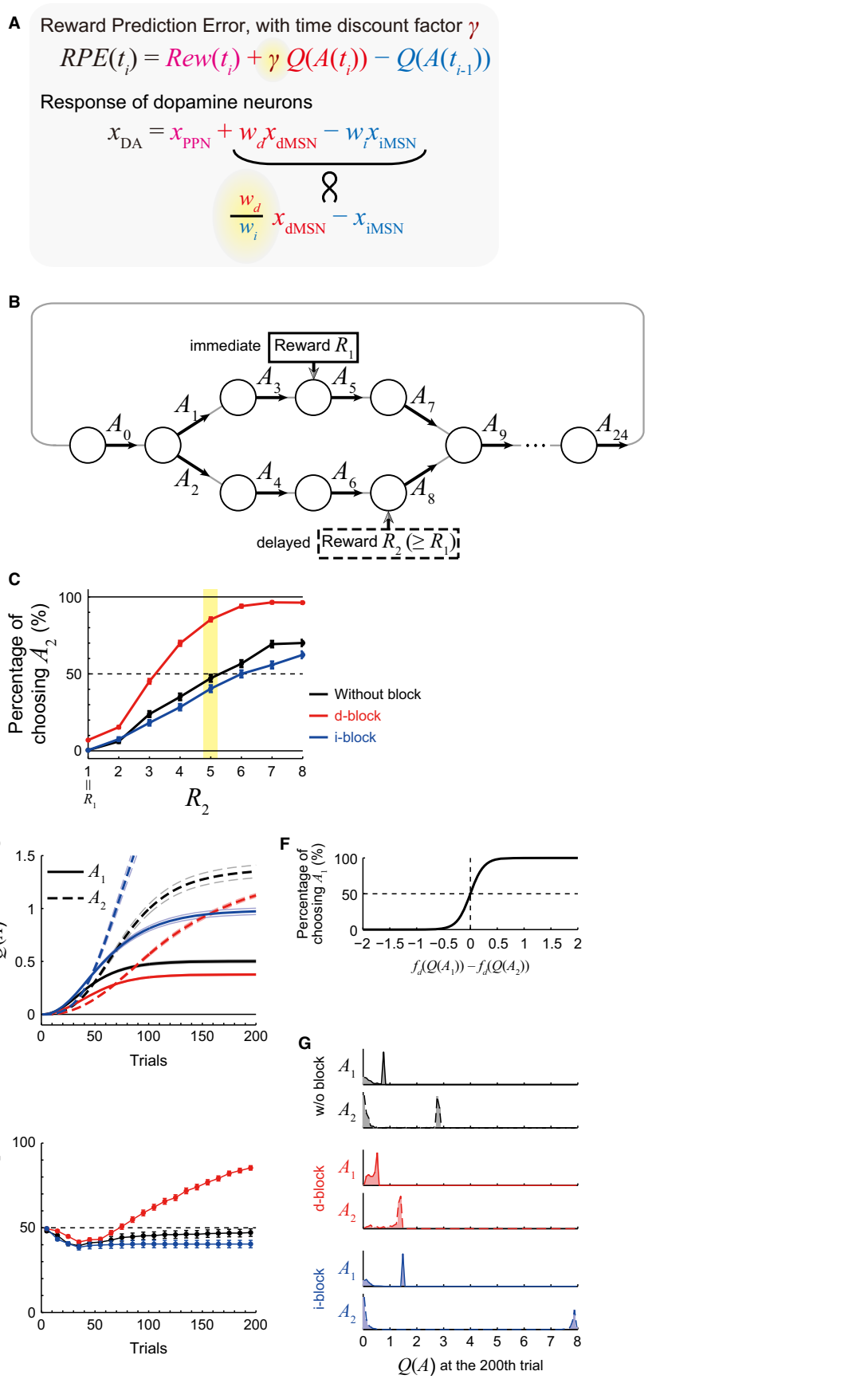

FIG. 7. Predictions of the CS-TD model on how the pathway-blockade (d-block/i-block) affects the subject's time preference. (A) According to the CS-TD model, the ratio of the strength of the direct pathway (w_d) over that of the indirect pathway (w_i) corresponds to the time discount factor (γ), which is included in the definition of RPE and represents the relative weight of upcoming (future) rewards over previous (past) rewards. (B) Simulated inter-temporal choice task. The amount of immediate reward (R_1) was set to be a constant ($R_1 = 1$), and eight cases with different amounts of delayed reward ($R_2 = 1, 2, \ldots$ or 8) were simulated (200 trials × 500 simulations for each case). (C–G) Simulation results. (C) Percentage of choosing delayed reward (R_2) in the 191–200th trials averaged across simulations in each of the eight cases with different amounts of delayed reward (horizontal axis). The lines and error bars represent the average and ± SEM, respectively, across simulations (the same is applied to E). The black, red, and blue colors indicate the control (without blockade), d-block, and i-block conditions, respectively (the same is applied to all of the following panels). (D) Trial-by-trial development of the value (reward expectation) of A_1 [$Q(A_1)$, solid lines] and A_2 [$Q(A_2)$, dashed lines] averaged across simulations. The thick and thin lines indicate the average and ± SEM, respectively, across simulations. The case of $R_2 = 5R_1$ is shown (same for D, E, and G). (E) Changes of the across-simulation percentage of choosing delayed reward along with trials (with the bin size = 10 trials). (F) Probabilistic action selection depending on $Q(A_1)$ and $Q(A_2)$ assumed in the simulations shown in C–G. (G) Distribution of $Q(A_1)$ and $Q(A_2)$ at the 200th trial across simulations.

ative or positive shift of RPE. The initial learning and the early phase of reversal learning would basically be driven by positive RPE and negative RPE, respectively, and they are thus expected to be impaired by the negative shift and positive shift of RPE caused by d-block and i-block. In order to examine whether such conjectures are valid, we simulated the reversal learning task by using the CS-TD model. To do so, we have made a simplification of the task. Specifically, as the two subtasks, with different cue-target/action associations (see above and Fig. 2A), appear to be equivalent and not much inter-related, in the sense that two cues in front of the subject (i.e. at the opposite side of the start) are not overlapped between the two subtasks, it would be expected that initial learning, as well as learning after a contingency reversal, proceeds in parallel and with similar speeds for the two subtasks. Therefore, we considered a single task with a contingency reversal. Also, there were three possible directions (or four, including the start location) in the maze used in the experiment, but we considered a two-alternative choice task (Fig. 2C). We assumed that the state transitions of the task, such as the diagram shown in the top of Fig. 2C, are represented in the subject's cerebral cortex, in particular in the orbitofrontal cortex (Wilson *et al.*, 2014) that projects to the NAc. At each presumed 'time-step' (t_i), the subject is assumed to exist at one of the states (circles in Fig. 2C) and is taking an action $A(t_i)$; if there are two options (A_1 or A_2), the subject is choosing one of them depending on their values (Fig. 2C, bottom inset). We assumed that, at each time-step/state, calculation of the RPE and RPE-dependent update of the value of previous action, together with action selection and execution, is conducted in the corticobasal ganglia–midbrain neural circuit according to the CS-TD model. We then simulated the task under three conditions: control (without block), d-block, and i-block.

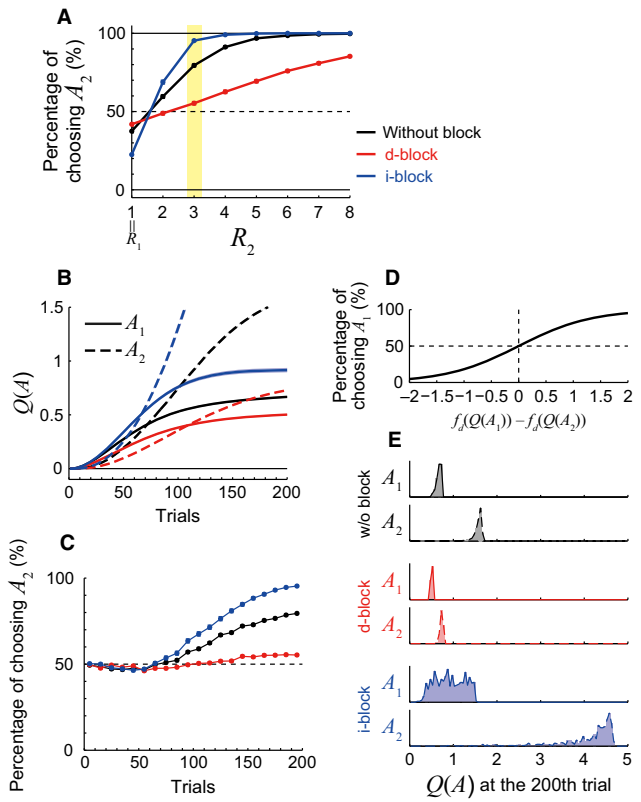

FIG. 8. Results of the simulations of the inter-temporal choice task with a larger degree of exploration upon action selection (as shown in D). Configurations are the same as those in Fig. 7C–G, except that cases of $R_2 = 3R_1$, rather than cases of $R_2 = 5R_1$, are shown in (B), (C), and (E).

Figure 3A shows the simulation results. As shown in the figure, our model can indeed reproduce the experimentally observed selective impairment of initial learning and learning after contingency reversal by d-block and i-block, respectively (Yawata *et al.*, 2012; Fig. 2B). The top panel of Fig. 3B shows the simulated RPE generated at S_2 in Fig. 2C (when action A_3 is being taken), which is used to update the predicted value of action A_1. As we conjectured in the above, d-block and i-block shifted RPE negatively and positively, respectively, during the initial learning and the early phase of reversed learning [notably, RPE at S_2 does not become negative immediately after the contingency reversal, whereas RPE at S_4 does (bottom panel of Fig. 3B)]. Figure 3C shows the number of trials needed to reach the performance criterion (≥ 95% rewarded choice in the last 20 trials) before (top panel) or after (bottom panel) the contingency reversal in the control (left bars), d-block (middle bars), and i-block (right bars) cases. Whether the criterion was reached or not was evaluated at every 10 trials in each session, and the error bars represent the mean ± SD across simulations. As shown in the top panel of Fig. 3C, the number of trials needed to reach the criterion in the initial learning was increased by d-block but not by i-block. However, the bottom panel of Fig. 3C indicates that the number of trials needed to reach the criterion after the contingency reversal was increased not only by i-block but also by d-block. This is because d-block did not impair the early phase of reversal learning but did impair the late phase (Fig. 3A), similar to the result of the experiment (Fig. 2B).

There is an apparent deviation of our simulation results (Fig. 3A) from the experiments (Fig. 2B) showing that switching of choice after the contingency reversal is rather slow. Increasing the learning rate in the model (from 0.05 to 0.1) resulted in faster learning (Fig. 3D), but not specifically at the moment after the contingency reversal. In fact, it has been suggested in human experiments (Behrens *et al.*, 2007) that the learning rate can be adaptively and dynamically changed depending on the volatility of the reward envi-

ronment. Therefore, we next incorporated adaptive modulation of the learning rate into our model. Specifically, we assumed that the learning rate was initially 0.05, and increased to a larger value (0.2) whenever the subjects experienced four consecutive rewarded trials followed by two consecutive unrewarded trials [modeling a change in the reward environment from the normal (i.e. mostly rewarded) pattern], and then decayed exponentially (with the time constant 20 trials). Figure 4 shows the simulation results for the elaborated model with the learning-rate modulation, which are closer to the experimental results [we also varied the degree of choice exploration (i.e. the slope of the sigmoid function shown in the bottom inset of Fig. 2C), confirming that the main features are largely preserved (Fig. 4E and F)]. Importantly, in our model (both the original model and the elaborated model incorporating the adaptive modulation of the learning rate), iMSNs always operate concomitantly with dMSNs, except for a time delay. Indeed, as shown in Fig. 5, in the first session, the population activity of iMSNs in the trials in which the action leading to reward (A_1) is chosen gradually increases along with trials, similar to the population activity of dMSNs (Fig. 5A and B, left and middle panels). Also, notably, although there is a time delay between dMSNs and iMSNs, the time-courses of their population activities are largely overlapped. Another key point is that the activity of dMSNs is enhanced by i-block, especially in the early phase of the second session (right panels). These are testable predictions of the CS-TD model.

As we have shown, the CS-TD model can account for how i-block critically impairs learning after reversal, which could be regarded as avoidance learning from a relatively aversive outcome (i.e. omission of reward). However, it has been shown that i-block also critically impairs avoidance learning from an absolutely aversive outcome, i.e. punishment. Specifically, it has been shown (Hikida *et al.*, 2010) that, if mice received electric footshock when entering from a light chamber into a preferred dark chamber, they showed marked delay in stepping into the dark chamber the next time, but this effect was significantly attenuated in animals with i-block but not in those with d-block. In order to test if the CS-TD model can also account for such results, we considered a similar setting to the above simulation of the reversal learning task, but this time introducing punishment, modeled by negative reward (-1), after 200 trials of initial appetitive reward ($+0.5$) learning (which was assumed so that the preference for the dark chamber appears) instead of the contingency reversal of positive reward (Fig. 6A). We conducted simulations with the same model parameters as used before (in Fig. 3A–C), and recorded how many trials were needed to set out to avoid the punishment (i.e. to choose the punishment-free option) for the first time after punishment was introduced, asking how it was affected by d-block and i-block. The left panel of Fig. 6B shows the results. As shown in Fig. 6B, i-block, but not d-block, increased the number of trials required for the acquisition of avoidance. This result appears to be in line with the experimental suggestion. More precisely, however, footshock was given not repeatedly but only once in the experiments (Hikida *et al.*, 2010), and so what matters would be whether the subject shows avoidance in the single trial following the punishment. In our simulation, the percentage of avoidance in this trial was rather low in any condition (right panel of Fig. 6B). However, this is certainly because the learning rate was kept constant at a small value (0.05), and so we conducted an additional set of simulations assuming that punishment induces an increase of the learning rate (0.05–0.5). As a result, the percentage of avoidance in the trial following the punishment was raised to around 50% in the control condition (Fig. 6C). Crucially, i-block, but not d-block, impaired such an avoidance developed by one-trial footshock (Fig. 6C), consistent with the experimental result (Hikida *et al.*, 2010).

In fact, the specific impairment of avoidance learning by i-block in the simulations shown above is considered to come, in a significant part, from the effect of i-block on the initial appetitive learning phase. Specifically, during the appetitive learning phase, the i-block-induced positive shift of RPE causes an exaggerated positive value of the action leading to reward (Fig. 6D), which is more difficult to cancel/reverse by negative RPE induced by reward omission or punishment. Although we think that this is an interesting and conceivable mechanism, it seems unclear whether or how well it can be applied to the case with presumably innate, rather than learned, preference, such as the preference for a dark chamber in the experiment considered above (Hikida *et al.*, 2010). Therefore, we conducted another set of simulations, in which appetitive reward learning was not assumed but instead action values were set *a priori*, with the actions corresponding to the dark chamber having larger values, regardless of the existence/type of pathway blockade (Fig. 6E). The learning rate was assumed to be initially 0 and increased to 0.5 upon receiving punishment (-1). Figure 6F shows the simulation results for the percentage of avoidance in the trial following the punishment. As shown in Fig. 6F, avoidance was impaired in the case of i-block compared with the control and d-block cases. This impairment is considered to occur purely because punishment-induced negative RPE was degraded (i.e. shifted positively) by i-block.

Predictions of the corticostriatal temporal difference model for the neural basis of time preference

Next we asked whether the CS-TD model can make predictions regarding behavior, in addition to the predictions about neural activity described before (Fig. 5). According to the CS-TD model, the relative strength/efficacy of the direct pathway over the indirect pathway corresponds to the 'time discount factor', which represents the relative weight of upcoming (future) rewards over previous (past) rewards (Fig. 7A; Morita *et al.*, 2012). Specifically, the weaker the direct pathway (indirect pathway) becomes, the smaller (larger) the time discount factor should become, i.e. the more severe (milder) the temporal discount of future rewards should become. Therefore, we conjectured that d-block or i-block, as considered in the above, would make the subject more strongly prefer small immediate reward or large future reward, respectively. We examined whether this conjecture actually holds in a simulated inter-temporal choice task, in which one action (A_1) is associated with immediate reward, whereas the other action (A_2) is associated with delayed reward (Fig. 7B). We simulated the subject's behavior and neural activity in this task for eight cases with different amounts of delayed reward, by using the CS-TD model (the original model with a constant learning rate of 0.05 used for the simulations shown in Fig. 3A–C) in a similar fashion to the simulation of reversal learning task in the above.

Figure 7C shows the simulation results, with the black, red, and blue lines indicating the percentage of choosing the delayed reward (A_2) in the 191–200th trials averaged across 500 independent simulations in the control, d-block, and i-block conditions, respectively, in each of the cases with different amounts of delayed reward (horizontal axis). As shown in Fig. 7C, d-block increases the percentage of choosing delayed larger reward, whereas i-block has the opposite effect. To our surprise, this result is exactly the opposite to what we expected in the above. Considering carefully, however, we realized that this result could also be intuitively understood. In the model,

development of the value (reward expectation) of action A_1 [denoted by $Q(A_1)$] precedes the development of the value of action A_2 [$Q(A_2)$], because A_1 is closer to reward than A_2 (Fig. 7D, showing the case with $R_2 = 5R_1$). d-block impairs this initial development of $Q(A_1)$ because it causes a negative shift of RPE (Fig. 7D, red solid line compared with black solid line). In contrast, i-block causes a positive shift of RPE and thereby boosts the development of $Q(A_1)$ (Fig. 7D, blue solid line). Such impaired/enhanced development of $Q(A_1)$ by d-block/i-block is considered to lead to a decrease/increase in the percentage of choosing A_1, and thus an increase/decrease in the percentage of choosing A_2, in the early phase before $Q(A_2)$ develops.

Figure 7E shows how the percentage of choosing A_2, averaged across simulations, changes along with trials (again in the case with $R_2 = 5R_1$). As shown in Fig. 7E, in the case with d-block, the percentage initially decreases but then turns to an increase, reflecting the delayed development of $Q(A_2)$. In contrast, in the case with i-block, the percentage of choosing A_2 decreases, more prominently than in the case of d-block due to the enhanced development of $Q(A_1)$ mentioned above, and then remains there; it almost never turns to an increase. This is presumably because, in a large part of the simulation runs, $Q(A_1)$ becomes so large that A_1 comes to be chosen in almost all of the trials, and conversely, A_2 is rarely selected and thus its value cannot be properly updated. Indeed, according to the probabilistic action selection assumed in the model (Fig. 7F; the same function as the one shown in the bottom inset in Fig. 2C), once the difference $f_d(Q(A_1)) - f_d(Q(A_2))$, where f_d is the assumed neuronal input–output transformation function of dMSNs, exceeds a certain level, there is almost no chance for A_2 to be chosen. This is clearly reflected in the across-simulation distribution of $Q(A_2)$ at the 200th trial (Fig. 7G). In the control condition (black), there are two peaks, one near 0 and the other at a large value, and the near-0 peak becomes more dominant in the condition with i-block (blue), whereas it almost disappears in the d-block case (red), explaining the results of Fig. 7C.

Now, what happens if the degree of exploration upon action selection is much larger, i.e. the slope of the sigmoid function of Fig. 7F is much less? It would be expected that A_2 can then be chosen with a certain probability in the early phase even in the condition with i-block so that $Q(A_2)$ can properly grow and eventually become larger than $Q(A_1)$. Moreover, given such proper development of action values in all of the conditions, it is expected that eventually (i.e. after many trials) delayed larger reward is selected more frequently in the condition with i-block, and less frequently in the condition with d-block, than in the control condition without blockade as we originally conjectured, because of the expected effects of d-block/i-block on the time discount factor discussed above. We conducted simulations with a larger degree of exploration (Fig. 8D), and confirmed that such expectations indeed hold. Specifically, $Q(A_2)$ turns into an increase even in the condition with i-block (Fig. 8C), and eventually A_2 becomes chosen more frequently in the i-block case, and less frequently in the d-block case, than in the control case (Fig. 8A), opposite to the cases with the smaller degree of exploration (Fig. 7C).

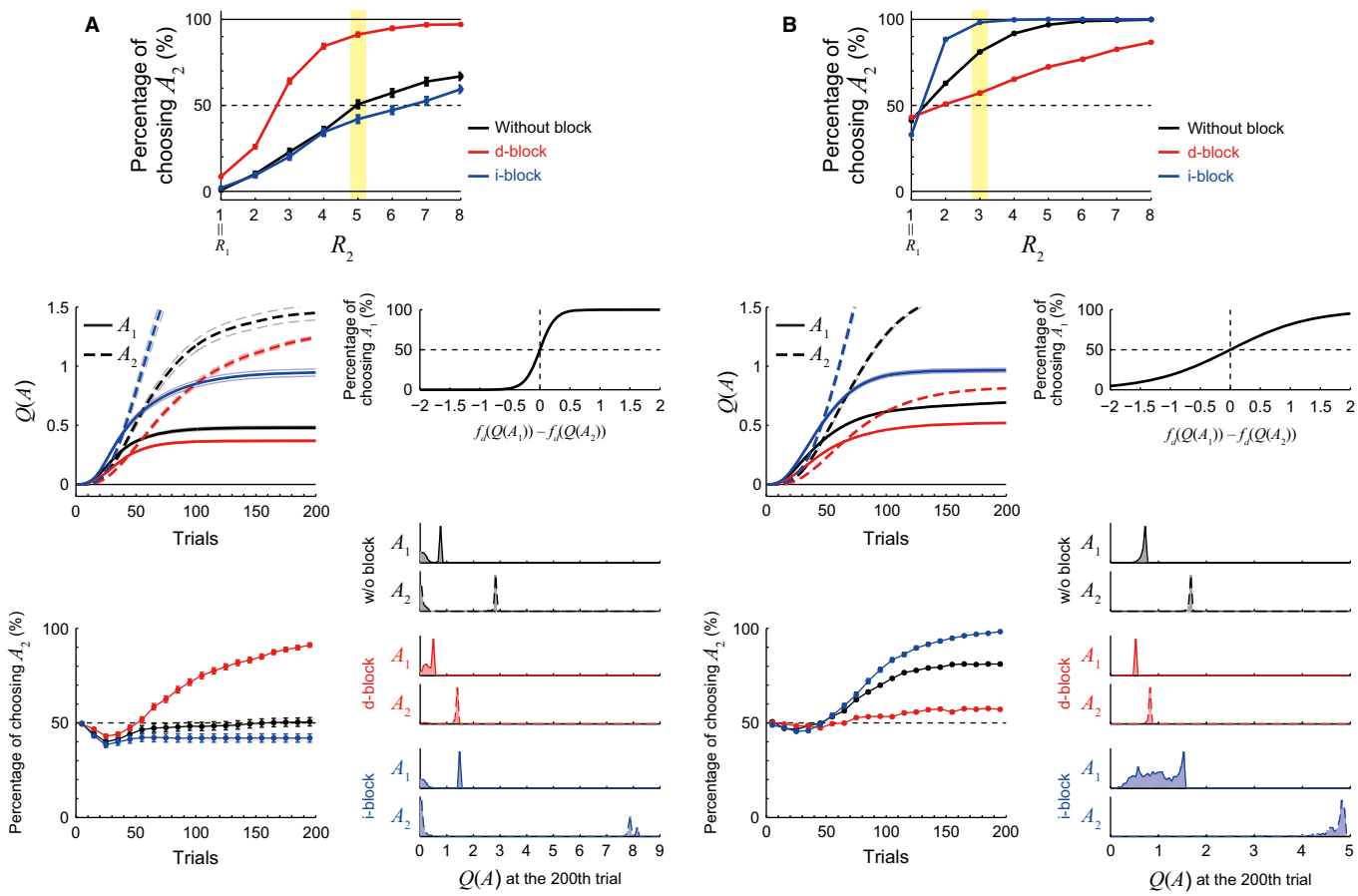

FIG. 9. Results of the simulations of the inter-temporal choice task with the elaborated model incorporating reward history-dependent adaptive modulation of the learning rate. A and B correspond to Figs 7C–G and 8, respectively.

In the simulations shown in Figs 7 and 8, we used the original model with a constant learning rate of 0.05 used for the simulations shown in Fig. 3A–C. We also conducted simulations by using the elaborated model with reward history-dependent adaptive modulation of the learning rate, which was used for the simulations shown in Fig. 4A–D. The results were largely similar, as shown in Fig. 9 (Fig. 9A and B correspond to Figs 7C–G and 8, respectively).

Discussion

There are two widely appreciated notions regarding the involvement of the corticobasal ganglia system in value-based learning: (i) the direct and indirect pathways of the basal ganglia (BG) control appetitive (Go) and aversive (No-Go) learning, respectively; and (ii) the midbrain dopamine neurons calculate RPE. The relationship between these two, however, has remained elusive, as described in the Introduction. In this work, we have explored an integrated mechanistic account for the two notions. Specifically, we examined whether a recently proposed model of the mechanism of RPE calculation, the CS-TD model, can also account for experimental results that suggest specialization of the direct and indirect pathways for appetitive and aversive learning. Through simulations of the reversal learning and avoidance learning tasks with pathway blockade, we have successfully addressed this issue, and also provided testable predictions on neural activity. In fact, the CS-TD model could also explain other results suggesting the specialization of the pathways, in particular seeking/avoidance of optogenetic self-stimulation of dMSNs/iMSNs (Kravitz *et al.*, 2012; see Morita *et al.*, 2013). We then asked if the CS-TD model can also provide predictions regarding behavior, and derived such predictions for time preference through simulation of inter-temporal choice. In the following, we discuss our model's plausibility, testable predictions, and implications for neural mechanisms of addiction.

Plausibility of the corticostriatal temporal difference model

In terms of the plausibility of the model's assumptions, the one regarding the dopamine-dependent modification of corticostriatal connection strengths would be the most controversial. There has been a significant amount of experimental results showing that dopamine induces or significantly modulates the plasticity of corticostriatal synapses (Reynolds *et al.*, 2001; Shen *et al.*, 2008; Yagishita *et al.*, 2014), and we have made the assumption with these results in mind, in a broad sense. However, it has actually been indicated that synapses on dMSNs [expressing dopamine D1 receptors (D1Rs)] and those on iMSNs [expressing dopamine D2 receptors (D2Rs)] entail opposite directions of plasticity. In particular, in spike-timing-dependent plasticity experiments with positive timing (pre-spike followed by post-spike), the potentiation of dMSN/D1 synapses and iMSN/D2 synapses was impaired and switched into depression by the application of D1R antagonist and D2R agonist, respectively (Shen *et al.*, 2008). Seemingly in line with these results, it has been proposed (Frank *et al.*, 2004; Hong & Hikosaka, 2011; Yawata *et al.*, 2012) that dMSN/D1 synapses and iMSN/D2 synapses are potentiated by dopamine increase and decrease, respectively. In contrast, our assumption posits that both dMSN/D1 and iMSN/D2 connections are strengthened by dopamine increase.

There are at least four factors that, we think, permit us to make such an assumption. First, most previous studies regarding corticostriatal plasticity did not distinguish CCS and CPn/PT inputs. Synapses made by CCS cells and those made by CPn/PT cells might actually have different plasticity properties. Also, induction of the same direction of plasticity in dMSNs and iMSNs might require different temporal patterns of inputs, in particular, rapidly developed CCS inputs and slowly developed (and sustained) CPn/PT inputs, respectively. Second, as D2Rs have a higher dopamine affinity than D1Rs (Richfield *et al.*, 1989), bath application of a high concentration of dopamine could mask changes in the degree of activation of D2Rs (but not those of D1Rs; Fig. 10). If such changes, rather than the tonic level of receptor activation, cause potentiation, potentiation of iMSN/D2 synapses would be blocked by a high concentration of dopamine. Third, most previous studies examined plastic changes of the AMPA response to a brief non-repetitive input, but synapse-type-dependent short-term facilitation might actually occur (Morita, 2014) and *N*-methyl-D-aspartate (NMDA) currents might also significantly contribute to corticostriatal transmission. Fourth, a recent article (Keeler *et al.*, 2014) has proposed the intriguing possibility that iMSNs positively encode rewards through the dopamine-induced internalization of D2Rs, which presumably reduces suppression of the neuronal responsiveness by tonic D2R activation [the entire proposal of this article (Keeler *et al.*, 2014) is also interesting; their proposal shares the feature that dMSNs and iMSNs encode rewards positively (rather than positively and negatively) with our CS-TD model, whereas differential activation of dMSNs/iMSNs by CCS and CPn/PT cells is not considered in their proposal]. Also, another recent study (Gurney *et al.*, 2015) has presented a detailed model of corticostriatal plasticity, in which the direction of change for appetitive learning is the same for dMSNs and iMSNs. Given these four factors, and also additional points discussed in our previous article (Morita *et al.*, 2013), we think that our assumption regarding plasticity can be valid, although it is surely a critical assumption of our model.

Along with the plasticity issue, another crucial issue is the selectivity in the connections between the different populations of the cortical and striatal cells. As described before, although there exists an apparent contradiction between anatomical results (Lei *et al.*, 2004; Reiner *et al.*, 2010) and physiological results (Ballion *et al.*, 2008; Kress *et al.*, 2013) regarding whether or not there exist CCS→dMSN and CPn/PT→iMSN connection preferences, it could be resolved, and the assumptions of the CS-TD model could still be valid, if effects of short-term plasticity are taken into account (Morita, 2014) and/or inputs to spines and those to shafts are distinguished (Deng *et al.*, 2015). In fact, there is another, totally different, possibility, apart from the dichotomy between CCS and CPn/PT cells. Specifically, a recent study (Wall *et al.*, 2013) has shown that the sensory cortex and motor cortex preferentially project to dMSNs and iMSNs, respectively. As information would naturally flow from the sensory cortex to the motor cortex, such preferences could result in the dMSNs and iMSNs representing the current and previous values, just as assumed in the CS-TD model but for a different reason.

In addition to the issues discussed so far, there are also important issues regarding the anatomy and physiology of the corticobasal ganglia circuits, including the involvement of other pathways such as the intrastriatal (Calabresi *et al.*, 2014), corticosubthalamic (Nambu *et al.*, 2002), thalamostriatal (Smith *et al.*, 2009), pallidostriatal (Mallet *et al.*, 2012), and pallidocortical (Chen *et al.*, 2015; Saunders *et al.*, 2015) connections, the heterogeneity of cortical/basal-ganglia regions (O'Doherty *et al.*, 2004; Wickens *et al.*, 2007; Thorn *et al.*, 2010; Rushworth *et al.*, 2011; Hunt *et al.*, 2012; Wilson *et al.*, 2014) and dopamine neurons (Bromberg-Martin *et al.*, 2010; Henny *et al.*, 2012), or the effects of individual neuronal spikes on plasticity (Shen *et al.*, 2008) or dynamics (Bevan *et al.*, 2002), which were not considered in our model. Some of these issues have been considered in other models (Humphries *et al.*,

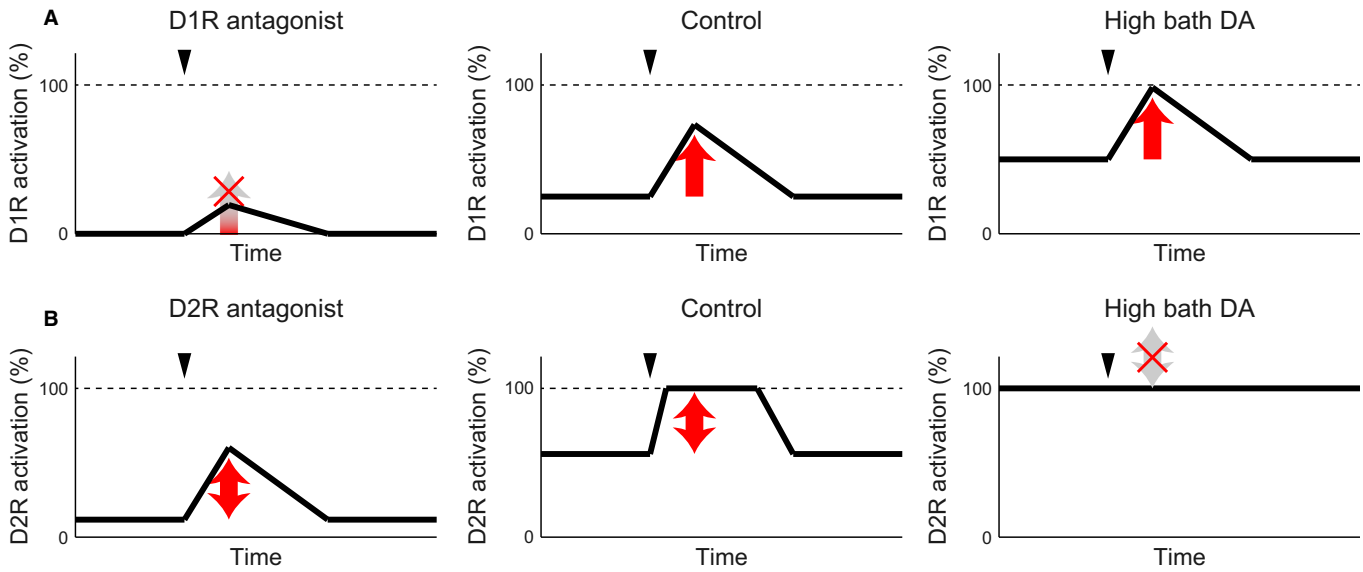

FIG. 10. Schematic illustrations of the presumed effects of bath application of dopamine (DA) receptor antagonist or high concentration of DA on the degree of DA receptor activation *in vitro*. The black arrowhead in each panel indicates stimulation of dopaminergic fibers in the slice, which causes a phasic release of DA. (A) Cases for D1Rs, which have relatively low DA affinity. Application of D1R antagonist (left panel) causes a reduction in both the baseline and the stimulation-induced phasic increase of the degree of D1R activation, compared with the control condition (middle panel). Application of a high concentration of DA (right panel) increases the baseline level of D1R activation, but saturation would not occur because of the low affinity. (B) Cases for D2Rs, which have relatively high DA affinity. Because of the high affinity, the baseline level of D2R activation in the control condition (middle panel) is higher than that of D1R activation. Stimulation-induced increase of the degree of D2R activation is also large, presumably causing transient saturation (middle panel); stimulation can induce a significant increase of the degree of D2R activation even with the presence of D2R antagonist (left panel). However, application of high concentration of DA or D2R agonist (right panel) causes saturation of D2Rs, and so stimulation cannot cause changes in the degree of D2R activation.

2006; Frank *et al.*, 2007; Potjans *et al.*, 2011; Collins & Frank, 2014; Schroll *et al.*, 2014; Gurney *et al.*, 2015). Not all of the issues have been considered at once, however, and also the other models have not considered the heterogeneity of corticostriatal inputs nor addressed how RPE is calculated, whereas our CS-TD model did so. Therefore, we think that the existing models, including ours, are complementary to each other, and a possible combination of them is expected to be explored. There are also many other unresolved computational issues, including how cortical representations of states or actions are formed (cf. Alexander & Brown, 2011; Wilson *et al.*, 2014), how timings and durations are represented (cf. Houk *et al.*, 1995; Joel *et al.*, 2002; Mauk & Buonomano, 2004; Daw *et al.*, 2006; Nakahara & Kaveri, 2010; Bernacchia *et al.*, 2011; Gershman *et al.*, 2014), and whether and how different reinforcement-learning algorithms, such as Q-learning, actor-critic, or SARSA, are implemented (cf. Joel *et al.*, 2002; Morris *et al.*, 2006; Niv *et al.*, 2006; Roesch *et al.*, 2007). The relationships between existing proposals on these issues and the CS-TD model are also expected to be explored in future studies.

Testable predictions of the corticostriatal temporal difference model

As shown in the Results, the CS-TD model predicts that dMSN and iMSN populations exhibit similar task-related activation with certain time differences, and also that the activity of both populations increases along with an increase of reward expectation (Fig. 5). A recent study (Isomura *et al.*, 2013) has shown that, in a reward-associated motor learning task, both dMSNs and (putative) iMSNs exhibit task-related activation, and the activity of both types of cells is mostly positively modulated by reward expectation. These results are potentially in accord with our prediction, although a significant time difference between the two cell types was not observed (possibly because of the limited number of examined cells). More direct test of the model's predictions about the activity of dMSNs and iMSNs is expected to be conducted by applying recently developed sophisticated methods (Cui *et al.*, 2014; Jin *et al.*, 2014) to value-based learning and choice tasks. Moreover, given that transgenic lines in which Cre-recombinase is selectively expressed in CCS or CPn/PT corticostriatal cell populations have recently been developed (Gerfen *et al.*, 2013) and have begun to be used for *in vivo* recording and manipulation (Li *et al.*, 2015), testing the CS-TD model by using these transgenic lines is also expected. Notably, however, the model posits that CCS and CPn/PT cells represent actions (or state-action pairs) rather than their learned values. Intriguingly, it has been shown for eyelid conditioning (Kalmbach *et al.*, 2009; Siegel *et al.*, 2012; Siegel & Mauk, 2013) that learning the association between temporally non-overlapped conditioned and unconditioned stimuli requires sustained activation of the corticopontine pathway, which presumably originates from CPn/PT cells. These results indicate that information carried by prefrontal sustained activity is not limited to explicit 'working memory' but more general. The assertion of our CS-TD model that the sustained activity of CPn/PT cells represents the previous action/state in value-based learning is in line with this, and suggests a wider role of prefrontal (CPn/PT) sustained activity in learning involving the BG and dopamine.

Crucially, the CS-TD model does not necessarily predict that only CPn/PT cells show activity in the 'delay period' in a task. Rather, the model predicts that, although both CCS cells and CPn/PT cells may show such activity, underlying mechanisms, as well as encoded information, should be different. As for the mechanisms, CCS delay activity should be caused by inputs from other layers, areas, or brain regions, whereas CPn/PT delay activity should be sustained by recurrent excitation within the population, presumably via NMDA

receptor stimulation (Wang, 1999) and/or short-term synaptic facilitation (Hempel *et al.*, 2000; Mongillo *et al.*, 2008). This (former) prediction is expected to be tested by local NMDA receptor blockade (Wang *et al.*, 2013). Regarding the encoded information, there should be a temporal difference, with CCS and CPn/PT cells encoding newer and older information, respectively. In particular, CCS and CPn/PT delay activity might encode prospective and retrospective working memory, respectively (cf. Fuster, 2000). This prediction is expected to be tested by recording specifically from CCS or CPn/PT cells (Li *et al.*, 2015) during tasks including a sequence of actions (e.g. Hikosaka *et al.*, 1995) or transformation of information (e.g. Takeda & Funahashi, 2002). Notably, however, single 'action' at the macroscopic level can be internally composed of a set of action elements, each of which is represented by a different subpopulation of CCS and CPn/PT cells. In that case, the temporal difference in the encoded information between CCS and CPn/PT cell subpopulations would not be large and sophisticated techniques would be required for detection.

In addition to the predictions regarding neural activity, the CS-TD model has also provided the prediction that manipulation of the strength of the direct or indirect pathway has bidirectional effects on the subject's time preference depending on her/his tendency for making exploratory over exploitative choices. There have been suggestions that the degree of exploration is controlled by noradrenaline (Usher *et al.*, 1999; Doya, 2002; Aston-Jones & Cohen, 2005), and these are now being experimentally examined by pharmacological manipulations of adrenergic receptors (Luksys *et al.*, 2009). Our predictions are thus expected to be tested by combining such experiments with the d-block/i-block (Hikida *et al.*, 2010; Yawata *et al.*, 2012) or other techniques for selective manipulation of the BG pathways (Kravitz *et al.*, 2010, 2012; Lobo *et al.*, 2010; Grueter *et al.*, 2013). It has been shown (Dembrow *et al.*, 2010), however, that CPn/PT cells and the commissural cells, which are partially overlapped with CCS cells, are differentially modulated by noradrenaline and acetylcholine. Therefore, the suggested noradrenergic modulation of the degree of exploration and the modulation of time preference by the balance of the BG pathways suggested by the present study would not be separate phenomena but are likely to be inter-related, and their overall circuit mechanisms are expected to be explored in the future.

Implications of the corticostriatal temporal difference model for neural mechanisms of addiction

Lastly, we would like to note that the CS-TD model provides an insight into the mechanisms of addiction to drugs that enhance the effects of dopamine in the striatum, such as cocaine (Fig. 11). Based on the notion that dopamine represents RPE (Montague *et al.*, 1996; Schultz *et al.*, 1997), it has been proposed (Redish, 2004) that the drug-induced enhancement of the effects of dopamine may correspond to an addition of a non-compensable positive term (i.e. a positive term that is never canceled out by prediction) to RPE, resulting in an unbounded increase of the values of states leading to drug receipt (Fig. 11B). We would like to add to this proposal that such drugs may also indirectly cause an unnatural positive shift of RPE via the BG pathways. Specifically, drug-induced enhancement of the effects of striatal (tonic) dopamine may cause up-regulation of the responsiveness of dMSNs through D1Rs and down-regulation of the responsiveness of iMSNs through D2Rs. According to the CS-TD model, such up-regulation and down-regulation of dMSNs and iMSNs, respectively, correspond to an increase of the 'upcoming value' term $[+\gamma Q(A(t_i))]$ and an attenuation of the negative 'previous value' term $[-Q(A(t_{i-1}))]$ in RPE (Fig. 11C), and thereby dually

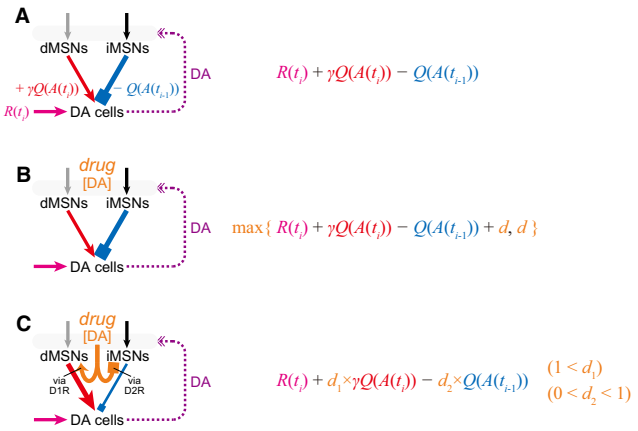

FIG. 11. Implication of the CS-TD model for neural mechanisms of drug addiction. (A) The presumed mechanism of RPE calculation in the dopamine (DA) neurons according to the CS-TD model. The schematic shown on the left is a simplification of a part of Fig. 1A. (B) Drug-induced enhancement of the effects of striatal DA has been proposed (Redish, 2004) to correspond to an addition of a non-compensable positive term (i.e. positive term that is never canceled out by prediction) to RPE (regardless of the mechanism of RPE calculation). (C) According to the CS-TD model, drugs may also indirectly cause an unnatural increase of RPE via the BG pathways. Drug-induced enhancement of the effects of striatal (tonic) DA may cause up-regulation of the responsiveness of dMSNs through D1Rs and down-regulation of the responsiveness of iMSNs through D2Rs. According to the CS-TD model, such up-regulation and down-regulation of dMSNs and iMSNs, respectively, correspond to an increase of the 'upcoming value' term $[+\gamma Q(A(t_i))]$ and an attenuation of the negative 'previous value' term $[-Q(A(t_{i-1}))]$ in RPE, dually resulting in a positive shift of RPE.

cause an unnatural positive shift of RPE. If this mechanism is indeed involved in the causes of addiction, enhancing the relative strength of the indirect pathway is expected to have an ameliorating effect. A recent finding that strengthening the indirect pathway of the NAc promotes resilience to compulsive use of cocaine (Bock *et al.*, 2013) could accord with this possibility. Notably, however, according to the CS-TD model, enhancing the relative strength of the indirect pathway could also lead to a smaller time discount factor, i.e. more severe discount of future values (Fig. 7A), which would rather be a risk factor for addiction as has been suggested (for a recent review, see Story *et al.*, 2014). Thus, the general prediction of the CS-TD model is that the balance between the BG pathways is a possible factor related to addiction, but its relation can be bidirectional, depending on the degree of exploration.

Acknowledgements

This work was supported by Grants-in-Aid for Scientific Research on Innovative Areas 'Prediction and Decision Making' (no. 26120710) and 'Mesoscopic Neurocircuitry' (no. 25115709) of The Ministry of Education, Science, Sports and Culture of Japan and the Strategic Japanese–German Cooperative Programme on 'Computational Neuroscience' (project title: neural circuit mechanisms of reinforcement learning) of the Japan Agency for Medical Research and Development (AMED) to K.M., and Core Research for Evolutional Science and Technology of JST and Grant-in-Aid for Scientific Research (nos. 25250005, 25123723, 15H01456) from the Japan Society for the Promotion of Science to Y. K. The authors have no conflict of interest to declare.

Abbreviations

BG, basal ganglia; CCS, crossed-corticostriatal; CPn/PT, corticopontine/pyramidal tract; CS-TD, corticostriatal temporal difference; D1R, dopamine D1 receptor; D2R, dopamine D2 receptor; dMSN, direct-pathway medium spiny neuron; GPe, external segment of the globus pallidus; GPi, internal segment

of the globus pallidus; iMSN, indirect-pathway medium spiny neuron; MSN, medium spiny neuron; NAc, nucleus accumbens; PPN, pedunculopontine nucleus; RPE, reward-prediction error; SNr, substantia nigra pars reticulata; STN, subthalamic nucleus.

References

Aggarwal, M., Hyland, B.I. & Wickens, J.R. (2012) Neural control of dopamine neurotransmission: implications for reinforcement learning. *Eur. J. Neurosci.*, **35**, 1115–1123.

Alexander, W.H. & Brown, J.W. (2011) Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.*, **14**, 1338–1344.

Aston-Jones, G. & Cohen, J.D. (2005) An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, **28**, 403–450.

Ballion, B., Mallet, N., Bézard, E., Lanciego, J.L. & Gonon, F. (2008) Intratelencephalic corticostriatal neurons equally excite striatonigral and striatopallidal neurons and their discharge activity is selectively reduced in experimental parkinsonism. *Eur. J. Neurosci.*, **27**, 2313–2321.

Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. (2007) Learning the value of information in an uncertain world. *Nat. Neurosci.*, **10**, 1214–1221.

Bernacchia, A., Seo, H., Lee, D. & Wang, X.-J. (2011) A reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.*, **14**, 366–372.

Bevan, M.D., Magill, P.J., Terman, D., Bolam, J.P. & Wilson, C.J. (2002) Move to the rhythm: oscillations in the subthalamic nucleus-external globus pallidus network. *Trends Neurosci.*, **25**, 525–531.

Bock, R., Shin, J.H., Kaplan, A.R., Dobi, A., Markey, E., Kramer, P.F., Gremel, C.M., Christensen, C.H., Adrover, M.F. & Alvarez, V.A. (2013) Strengthening the accumbal indirect pathway promotes resilience to compulsive cocaine use. *Nat. Neurosci.*, **16**, 632–638.

Bromberg-Martin, E.S., Matsumoto, M. & Hikosaka, O. (2010) Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, **68**, 815–834.

Calabresi, P., Picconi, B., Tozzi, A., Ghiglieri, V. & Di Filippo, M. (2014) Direct and indirect pathways of basal ganglia: a critical reappraisal. *Nat. Neurosci.*, **17**, 1022–1030.

Chen, M.C., Ferrari, L., Sacchet, M.D., Foland-Ross, L.C., Qiu, M.H., Gotlib, I.H., Fuller, P.M., Arrigoni, E. & Lu, J. (2015) Identification of a direct GABAergic pallidocortical pathway in rodents. *Eur. J. Neurosci.*, **41**, 748–759.

Collins, A.G. & 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.

Cowan, R.L. & Wilson, C.J. (1994) Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. *J. Neurophysiol.*, **71**, 17–32.

Crittenden, J.R. & Graybiel, A.M. (2011) Basal Ganglia disorders associated with imbalances in the striatal striosome and matrix compartments. *Front. Neuroanat.*, **5**, 59.

Cui, G., Jun, S.B., Jin, X., Luo, G., Pham, M.D., Lovinger, D.M., Vogel, S.S. & Costa, R.M. (2014) Deep brain optical measurements of cell type-specific neural activity in behaving mice. *Nat. Protoc.*, **9**, 1213–1228.

Daw, N.D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711.

Daw, N.D., Courville, A.C., Tourtezky, D.S. & Touretzky, D.S. (2006) Representation and timing in theories of the dopamine system. *Neural Comput.*, **18**, 1637–1677.

Dembrow, N.C., Chitwood, R.A. & Johnston, D. (2010) Projection-specific neuromodulation of medial prefrontal cortex neurons. *J. Neurosci.*, **30**, 16922–16937.

Deng, Y., Lanciego, J.L., Kerkerian-Le Goff, L., Coulon, P., Salin, P., Kachidian, P., Lei, W., Del Mar, N. & Reiner, A. (2015) Differential organization of cortical inputs to striatal projection neurons of the matrix compartment in rats. *Front. Syst. Neurosci.*, **9**, 51.

Ding, J., Peterson, J.D. & Surmeier, D.J. (2008) Corticostriatal and thalamostriatal synapses have distinctive properties. *J. Neurosci.*, **28**, 6483–6492.

Doya, K. (2002) Metalearning and neuromodulation. *Neural Networks*, **15**, 495–506.

Frank, M.J. (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *J. Cognitive Neurosci.*, **17**, 51–72.

Frank, M.J., Seeberger, L.C. & O'Reilly, R.C. (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, **306**, 1940–1943.

Frank, M.J., Samanta, J., Moustafa, A.A. & Sherman, S.J. (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, **318**, 1309–1312.

Fujiyama, F., Sohn, J., Nakano, T., Furuta, T., Nakamura, K.C., Matsuda, W. & Kaneko, T. (2011) Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector. *Eur. J. Neurosci.*, **33**, 668–677.

Fuster, J.M. (2000) Executive frontal functions. *Exp. Brain Res.*, **133**, 66–70.

Gerfen, C.R. (1984) The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature*, **311**, 461–464.

Gerfen, C.R. & Surmeier, D.J. (2011) Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci.*, **34**, 441–466.

Gerfen, C.R., Paletzki, R. & Heintz, N. (2013) GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron*, **80**, 1368–1383.

Gershman, S.J., Moustafa, A.A. & Ludvig, E.A. (2014) Time representation in reinforcement learning models of the basal ganglia. *Front. Comput. Neurosci.*, **7**, 194.

Gittis, A.H., Nelson, A.B., Thwin, M.T., Palop, J.J. & Kreitzer, A.C. (2010) Distinct roles of GABAergic interneurons in the regulation of striatal output pathways. *J. Neurosci.*, **30**, 2223–2234.

Graybiel, A.M. (2008) Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.*, **31**, 359–387.

Grueter, B.A., Robison, A.J., Neve, R.L., Nestler, E.J. & Malenka, R.C. (2013) ΔFosB differentially modulates nucleus accumbens direct and indirect pathway function. *Proc. Natl. Acad. Sci. USA*, **110**, 1923–1928.

Gurney, K.N., Lepora, N., Shah, A., Koene, A. & Redgrave, P. (2013) Action discovery and intrinsic motivation: a biologically constrained formalisation. In Baldassarre, G. & Mirolli, M. (Eds), *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer, Berlin Heidelberg, pp. 151–181.

Gurney, K.N., Humphries, M.D. & Redgrave, P. (2015) A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS Biol.*, **13**, e1002034.

Hart, A.S., Rutledge, R.B., Glimcher, P.W. & Phillips, P.E. (2014) Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J. Neurosci.*, **34**, 698–704.

Hazy, T.E., Frank, M.J. & O'Reilly, R.C. (2010) Neural mechanisms of acquired phasic dopamine responses in learning. *Neurosci. Biobehav. R.*, **34**, 701–720.

Hempel, C., Hartman, K., Wang, X., Turrigiano, G. & Nelson, S. (2000) Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.*, **83**, 3031–3041.

Henny, P., Brown, M.T., Northrop, A., Faunes, M., Ungless, M.A., Magill, P.J. & Bolam, J.P. (2012) Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons. *Nat. Neurosci.*, **15**, 613–619.

Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi, S. (2010) Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron*, **66**, 896–907.

Hikosaka, O., Rand, M.K., Miyachi, S. & Miyashita, K. (1995) Learning of sequential movements in the monkey: process of learning and retention of memory. *J. Neurophysiol.*, **74**, 1652–1661.

Hikosaka, O., Takikawa, Y. & Kawagoe, R. (2000) Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol. Rev.*, **80**, 953–978.

Hikosaka, O., Nakamura, K. & Nakahara, H. (2006) Basal ganglia orient eyes to reward. *J. Neurophysiol.*, **95**, 567–584.

Hirai, Y., Morishima, M., Karube, F. & Kawaguchi, Y. (2012) Specialized cortical subnetworks differentially connect frontal cortex to parahippocampal areas. *J. Neurosci.*, **32**, 1898–1913.

Hong, S. & Hikosaka, O. (2011) Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. *Front. Behav. Neurosci.*, **5**, 15.

Houk, J., Adams, J. & Barto, A. (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J.C., Davis, J.L. & Beiser, D.G. (Eds), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, USA.

Humphries, M.D., Stewart, R.D. & Gurney, K.N. (2006) A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J. Neurosci.*, **26**, 12921–12942.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F. & Behrens, T.E. (2012) Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.*, **15**, 470–476.

Isomura, Y., Takekawa, T., Harukuni, R., Handa, T., Aizawa, H., Takada, M. & Fukai, T. (2013) Reward-modulated motor information in identified striatum neurons. *J. Neurosci.*, **33**, 10209–10220.
Ito, M. & Doya, K. (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.*, **29**, 9861–9874.
Jin, X., Tecuapetla, F. & Costa, R.M. (2014) Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat. Neurosci.*, **17**, 423–430.
Jocham, G., Hunt, L.T., Near, J. & Behrens, T.E. (2012) A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nat. Neurosci.*, **15**, 960–961.
Joel, D., Niv, Y. & Ruppin, E. (2002) Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, **15**, 535–547.
Kalmbach, B.E., Ohyama, T., Kreider, J.C., Riusech, F. & Mauk, M.D. (2009) Interactions between prefrontal cortex and cerebellum revealed by trace eyelid conditioning. *Learn. Memory*, **16**, 86–95.
Keeler, J.F., Pretsell, D.O. & Robbins, T.W. (2014) Functional implications of dopamine D1 vs. D2 receptors: a 'prepare and select' model of the striatal direct vs. indirect pathways. *Neuroscience*, **282C**, 156–175.
Kim, H., Sul, J.H., Huh, N., Lee, D. & Jung, M.W. (2009) Role of striatum in updating values of chosen actions. *J. Neurosci.*, **29**, 14701–14712.
Kravitz, A.V., Freeze, B.S., Parker, P.R., Kay, K., Thwin, M.T., Deisseroth, K. & Kreitzer, A.C. (2010) Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature*, **466**, 622–626.
Kravitz, A.V., Tye, L.D. & Kreitzer, A.C. (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.*, **15**, 816–818.
Kreitzer, A.C. & Malenka, R.C. (2007) Endocannabinoid-mediated rescue of striatal LTD and motor deficits in Parkinson's disease models. *Nature*, **445**, 643–647.
Kress, G.J., Yamawaki, N., Wokosin, D.L., Wickersham, I.R., Shepherd, G.M. & Surmeier, D.J. (2013) Convergent cortical innervation of striatal projection neurons. *Nat. Neurosci.*, **16**, 665–667.
Lau, B. & Glimcher, P.W. (2008) Value representations in the primate striatum during matching behavior. *Neuron*, **58**, 451–463.
Lei, W., Jiao, Y., Del Mar, N. & Reiner, A. (2004) Evidence for differential cortical input to direct pathway versus indirect pathway striatal projection neurons in rats. *J. Neurosci.*, **24**, 8289–8299.
Li, N., Chen, T.W., Guo, Z.V., Gerfen, C.R. & Svoboda, K. (2015) A motor cortex circuit for motor planning and movement. *Nature*, **519**, 51–56.
Lobb, C.J., Wilson, C.J. & Paladini, C.A. (2011) High-frequency, short-latency disinhibition bursting of midbrain dopaminergic neurons. *J. Neurophysiol.*, **105**, 2501–2511.
Lobo, M.K., Covington, H.E., Chaudhury, D., Friedman, A.K., Sun, H., Damez-Werno, D., Dietz, D.M., Zaman, S., Koo, J.W., Kennedy, P.J., Mouzon, E., Mogri, M., Neve, R.L., Deisseroth, K., Han, M.H. & Nestler, E.J. (2010) Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science*, **330**, 385–390.
Luksys, G., Gerstner, W. & Sandi, C. (2009) Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. *Nat. Neurosci.*, **12**, 1180–1186.
Mallet, N., Le Moine, C., Charpier, S. & Gonon, F. (2005) Feedforward inhibition of projection neurons by fast-spiking GABA interneurons in the rat striatum in vivo. *J. Neurosci.*, **25**, 3857–3869.
Mallet, N., Micklem, B.R., Henny, P., Brown, M.T., Williams, C., Bolam, J.P., Nakamura, K.C. & Magill, P.J. (2012) Dichotomous organization of the external globus pallidus. *Neuron*, **74**, 1075–1086.
Mauk, M.D. & Buonomano, D.V. (2004) The neural basis of temporal processing. *Annu. Rev. Neurosci.*, **27**, 307–340.
Mink, J.W. (1996) The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.*, **50**, 381–425.
Mongillo, G., Barak, O. & Tsodyks, M. (2008) Synaptic theory of working memory. *Science*, **319**, 1543–1546.
Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947.
Morishima, M. & Kawaguchi, Y. (2006) Recurrent connection patterns of corticostriatal pyramidal cells in frontal cortex. *J. Neurosci.*, **26**, 4394–4405.
Morishima, M., Morita, K., Kubota, Y. & Kawaguchi, Y. (2011) Highly differentiated projection-specific cortical subnetworks. *J. Neurosci.*, **31**, 10380–10391.
Morita, K. (2014) Differential cortical activation of the striatal direct and indirect pathway cells: reconciling the anatomical and optogenetic results by using a computational method. *J. Neurophysiol.*, **112**, 120–146.
Morita, K., Morishima, M., Sakai, K. & Kawaguchi, Y. (2012) Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends Neurosci.*, **35**, 457–467.
Morita, K., Morishima, M., Sakai, K. & Kawaguchi, Y. (2013) Dopaminergic control of motivation and reinforcement learning: a closed-circuit account for reward-oriented behavior. *J. Neurosci.*, **33**, 8866–8890.
Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. (2006) Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.*, **9**, 1057–1063.
Moyer, J.T., Halterman, B.L., Finkel, L.H. & Wolf, J.A. (2014) Lateral and feedforward inhibition suppress asynchronous activity in a large, biophysically-detailed computational model of the striatal network. *Front. Comput. Neurosci.*, **8**, 152.
Nakahara, H. & Kaveri, S. (2010) Internal-time temporal difference model for neural value-based decision making. *Neural Comput.*, **22**, 3062–3106.
Nakanishi, S., Hikida, T. & Yawata, S. (2014) Distinct dopaminergic control of the direct and indirect pathways in reward-based and avoidance learning behaviors. *Neuroscience*, **282**, 49–59.
Nambu, A., Tokuno, H. & Takada, M. (2002) Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neurosci. Res.*, **43**, 111–117.
Niv, Y., Daw, N.D. & Dayan, P. (2006) Choice values. *Nat. Neurosci.*, **9**, 987–988.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R.J. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452–454.
Okada, K., Toyama, K., Inoue, Y., Isa, T. & Kobayashi, Y. (2009) Different pedunculopontine tegmental neurons signal predicted and actual task rewards. *J. Neurosci.*, **29**, 4858–4870.
Plenz, D. & Kitai, S.T. (1998) Up and down states in striatal medium spiny neurons simultaneously recorded with spontaneous activity in fast-spiking interneurons studied in cortex-striatum-substantia nigra organotypic cultures. *J. Neurosci.*, **18**, 266–283.
Potjans, W., Diesmann, M. & Morrison, A. (2011) An imperfect dopaminergic error signal can drive temporal-difference learning. *PLoS Comput. Biol.*, **7**, e1001133.
Redgrave, P., Prescott, T.J. & Gurney, K. (1999) The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, **89**, 1009–1023.
Redish, A.D. (2004) Addiction as a computational process gone awry. *Science*, **306**, 1944–1947.
Reiner, A., Hart, N.M., Lei, W. & Deng, Y. (2010) Corticostriatal projection neurons – dichotomous types and dichotomous functions. *Front. Neuroanat.*, **4**, 142.
Reynolds, J.N., Hyland, B.I. & Wickens, J.R. (2001) A cellular mechanism of reward-related learning. *Nature*, **413**, 67–70.
Richfield, E.K., Penney, J.B. & Young, A.B. (1989) Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. *Neuroscience*, **30**, 767–777.
Robbins, T.W. & Everitt, B.J. (1996) Neurobehavioural mechanisms of reward and motivation. *Curr. Opin. Neurobiol.*, **6**, 228–236.
Roesch, M.R., Calu, D.J. & Schoenbaum, G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.*, **10**, 1615–1624.
Rushworth, M.F., Noonan, M.P., Boorman, E.D., Walton, M.E. & Behrens, T.E. (2011) Frontal cortex and reward-guided learning and decision-making. *Neuron*, **70**, 1054–1069.
Samejima, K., Ueda, Y., Doya, K. & Kimura, M. (2005) Representation of action-specific reward values in the striatum. *Science*, **310**, 1337–1340.
Sato, M. & Hikosaka, O. (2002) Role of primate substantia nigra pars reticulata in reward-oriented saccadic eye movement. *J. Neurosci.*, **22**, 2363–2373.
Saunders, A., Oldenburg, I.A., Berezovskii, V.K., Johnson, C.A., Kingery, N.D., Elliott, H.L., Xie, T., Gerfen, C.R. & Sabatini, B.L. (2015) A direct GABAergic output from the basal ganglia to frontal cortex. *Nature*, **521**, 85–89.
Schroll, H., Vitay, J. & Hamker, F.H. (2014) Dysfunctional and compensatory synaptic plasticity in Parkinson's disease. *Eur. J. Neurosci.*, **39**, 688–702.
Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.

Seo, M., Lee, E. & Averbeck, B.B. (2012) Action selection and action value in frontal-striatal circuits. *Neuron*, **74**, 947–960.
Shen, W., Flajolet, M., Greengard, P. & Surmeier, D.J. (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, **321**, 848–851.
Shepherd, G.M. (2013) Corticostriatal connectivity and its role in disease. *Nat. Rev. Neurosci.*, **14**, 278–291.
Siegel, J.J. & Mauk, M.D. (2013) Persistent activity in prefrontal cortex during trace eyelid conditioning: dissociating responses that reflect cerebellar output from those that do not. *J. Neurosci.*, **33**, 15272–15284.
Siegel, J.J., Kalmbach, B., Chitwood, R.A. & Mauk, M.D. (2012) Persistent activity in a cortical-to-subcortical circuit: bridging the temporal gap in trace eyelid conditioning. *J. Neurophysiol.*, **107**, 50–64.
Smith, Y., Raju, D., Nanda, B., Pare, J.F., Galvan, A. & Wichmann, T. (2009) The thalamostriatal systems: anatomical and functional organization in normal and parkinsonian states. *Brain. Res. Bull.*, **78**, 60–68.
Smith, R.J., Lobo, M.K., Spencer, S. & Kalivas, P.W. (2013) Cocaine-induced adaptations in D1 and D2 accumbens projection neurons (a dichotomy not necessarily synonymous with direct and indirect pathways). *Curr. Opin. Neurobiol.*, **23**, 546–552.
Soltani, A. & Wang, X.-J. (2008) From biophysics to cognition: reward-dependent adaptive choice behavior. *Curr. Opin. Neurobiol.*, **18**, 209–216.
Story, G.W., Vlaev, I., Seymour, B., Darzi, A. & Dolan, R.J. (2014) Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective. *Front. Behav. Neurosci.*, **8**, 76.
Takeda, K. & Funahashi, S. (2002) Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. *J. Neurophysiol.*, **87**, 567–588.
Tepper, J.M. & Lee, C.R. (2007) GABAergic control of substantia nigra dopaminergic neurons. *Prog. Brain Res.*, **160**, 189–208.
Tepper, J.M., Martin, L.P. & Anderson, D.R. (1995) GABAA receptor-mediated inhibition of rat substantia nigra dopaminergic neurons by pars reticulata projection neurons. *J. Neurosci.*, **15**, 3092–3103.
Thorn, C.A., Atallah, H., Howe, M. & Graybiel, A.M. (2010) Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, **66**, 781–795.
Ungless, M.A., Magill, P.J. & Bolam, J.P. (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, **303**, 2040–2042.
Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. (1999) The role of locus coeruleus in the regulation of cognitive performance. *Science*, **283**, 549–554.
Wall, N.R., De La Parra, M., Callaway, E.M. & Kreitzer, A.C. (2013) Differential innervation of direct- and indirect-pathway striatal projection neurons. *Neuron*, **79**, 347–360.
Wang, X.-J. (1999) Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.*, **19**, 9587–9603.
Wang, X.-J. (2002) Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, **36**, 955–968.
Wang, M., Yang, Y., Wang, C.J., Gamo, N.J., Jin, L.E., Mazer, J.A., Morrison, J.H., Wang, X.J. & Arnsten, A.F. (2013) NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron*, **77**, 736–749.
Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A. & Uchida, N. (2012) Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, **74**, 858–873.
Watkins, C. (1989) *Learning from Delayed Rewards*. PhD Thesis. University of Cambridge, Cambridge.
Wickens, J.R., Budd, C.S., Hyland, B.I. & Arbuthnott, G.W. (2007) Striatal contributions to reward and decision making: making sense of regional variations in a reiterated processing matrix. *Ann. NY Acad. Sci.*, **1104**, 192–212.
Wilson, R.C., Takahashi, Y.K., Schoenbaum, G. & Niv, Y. (2014) Orbitofrontal cortex as a cognitive map of task space. *Neuron*, **81**, 267–279.
Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C., Urakubo, H., Ishii, S. & Kasai, H. (2014) A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, **345**, 1616–1620.
Yawata, S., Yamaguchi, T., Danjo, T., Hikida, T. & Nakanishi, S. (2012) Pathway-specific control of reward learning and its flexibility via selective dopamine receptors in the nucleus accumbens. *Proc. Natl. Acad. Sci. USA*, **109**, 12764–12769.