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

Kenji Morita¹ and Yasuo Kawaguchi²,³,⁴
¹Physical and Health Education, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
²Division of Cerebral Circuitry, National Institute for Physiological Sciences, Okazaki, Japan
³Department of Physiological Sciences, SOKENDAI (The Graduate University for Advanced Studies), Okazaki, Japan
⁴Japan 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), 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; Nakamishi 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 learning 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.
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,
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 expectancy) 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

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

\[
x_{\text{iMSN}}(t_i) = f_i(Q(A(t_{i-1}))) \tag{2}\]

\( x_{\text{dMSN}}(t_i) \) and \( x_{\text{iMSN}}(t_i) \) 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 \( z > 0 \)) in the case without blockade of BG pathway. Notably, we hypothesized that different dMSN or iMSN subpopulations correspond to different actions, but the variable \( x_{\text{dMSN}} \) or \( x_{\text{iMSN}} \) 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:

\[
x_{\text{dMSN}}(t_i) = \max\{f_d(Q(A_1)), f_d(Q(A_2))\} \tag{3}\]

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

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

\( x_{\text{PPN}}(t_i) \) 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:

\[
x_{\text{DA}}(t_i) = x_{\text{PPN}}(t_i) + \gamma x_{\text{dMSN}}(t_i) - x_{\text{iMSN}}(t_i) \tag{5}\]

where \( x_{\text{DA}}(t_i) \) represents the response of dopamine neurons (or resulting dopamine release) compared with its baseline level \( x_{\text{DA}}(t_i) \) 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 \( \gamma \) 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. \( \gamma \) was set to 0.75 in the present work. The terms \( ^{+} \! x_{\text{dMSN}}(t_i)^{+} \) and \( ^{-} \! x_{\text{iMSN}}(t_i)^{-} \) were assumed to come through the dMSNs→nTn segment of the globus pallidus (GP)/SNr→ventral tegmental area/substrata nigra pars compacta (dis-inhibition, net positive) and the iMSNs→external segment of the globus pallidus (GP)→[subthalamic nucleus (STN)]→GP/SNr→ventral tegmental area (VTA)/substrata nigra pars compacta (dis-inhibition, net negative).
pars compacta (SNc) (dis-dis-inhibition, net negative), respectively. As such, the terms ‘+ \gamma x_{\text{MSN}}(t_i)' and ‘- x_{\text{MSN}}(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 these inputs, which are assumed to receive information about the values 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 ‘- \gamma x_{\text{MSN}}(t_i)' if the striosomal neurons receive preferential inputs from CPu/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 CPu/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 over 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 CPu/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:

\begin{align}
\text{Prob}(A_1) &= 1 / (1 + \exp(-f_i(Q(A_1) - f_i(Q(A_2)))/\epsilon)), \quad (6) \\
\text{Prob}(A_2) &= 1 / (1 + \exp(-f_i(Q(A_2) - f_i(Q(A_1)))/\epsilon)) \\
&= 1 - \text{Prob}(A_1) \quad (7)
\end{align}

where \text{Prob}(A_1) and \text{Prob}(A_2) represent the probability of choosing action A_1 and A_2, respectively (bottom inset in Fig. 2C). \epsilon is a parameter determining the ‘flatness’ of the sigmoidal (soft-max) function, and it represents the degree of exploration (over exploitation) upon action selection. \epsilon 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 CPu/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, 2006). K. Morita and Y. Kawaguchi

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_i is the assumed neuronal input-output transformation function of dMSNs. *P < 0.05, **P < 0.01, ***P < 0.001, respectively.

© 2015 The Authors. European Journal of Neuroscience published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.
European Journal of Neuroscience, 42, 2003–2021
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:

$$Q(A(t_{i-1})) \rightarrow Q(A(t_{i-1})) + \alpha_{DA}(t_i)$$  \hspace{1cm} (8)$$

where $\alpha$ 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 $\alpha$ 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:
\[ a = 0.05 + 0.15 \exp(-N_{ap} - 1)/20 \]  

(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 $S_3$ (receiving reward and taking $A_3$) in Fig. 2C, respectively. (B) Results of the simulations with the elaborated model incorporating reward history-dependent state transitions (Daw et al., 2010); and (iii) an inter-temporal choice task, which is a virtual task intended to model avoidance of electric footshock (Hikida et al., 2012) (see the Results for details about the rationale of this is that presumably the existence of the reversal many times). Also, we assumed that the corticobasal ganglia circuit according to 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 d- or i-MSNs. For all of the tasks, we assumed that a diagram of action-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).
A 201st trial
Reward
+0.5

0 1 2 3 4 5 6 7 8 9

A 24

B

Number of trials needed to choose the punishment-free action

Without block d-block i-block

Percentage of avoidance in the trial just after receiving punishment (%)

Without block d-block i-block

C

E

Punishment

Without block d-block i-block

Value : 1
Value : 1.5

0 50 100

Trials

D

Percentage of avoidance in the trial just after receiving punishment (%)

0 50 100

Without block d-block i-block

0

Trials

© 2015 The Authors. European Journal of Neuroscience published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

European Journal of Neuroscience, 42, 2003–2021

2010 K. Morita and Y. Kawaguchi
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. \( f_d(z) = 0 \) (if \( z \leq 0 \)) or \( f_i(z) = 0 \) (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 S1 if the subject chooses A1 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 A1 in the last 20 trials became \( \geq 95\% \), the action-reward contingency was assumed to be reversed so that reward is obtained at state S2 if the subject chooses A2 [‘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 \( \sim \) 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-
A. Reward Prediction Error, with time discount factor $\gamma$

\[ RPE(t) = Rew(t) + \gamma Q(A(t)) - Q(A(t_{t+1})) \]

Response of dopamine neurons

\[ x_{DA} = x_{PPN} + w_d x_{dMSN} - w_i x_{iMSN} \]

B. Immediate Reward $R_1$

\[ A_0 \rightarrow A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow A_6 \rightarrow A_7 \rightarrow A_8 \rightarrow \ldots \]

Delayed Reward $R_2 (\geq R_1)$

C. Percentage of choosing $A_2 (%)$

Without block

\[ -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 \]

0 50 100

1 2 3 4 5 6 7 8

D. $Q(A)$

\[ f(Q(A)) = f(Q(A)) \]

Trials

0 50 100 150 200

0 0.5 1 1.5

E. Percentage of choosing $A_2 (%)$

Without block

\[ 0 50 100 \]

Trials

0 50 100 150 200

0 50 100

F. Percentage of choosing $A_1 (%)$

\[ -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 \]

Without block

\[ 0 50 100 \]

Trials

0 50 100

G. $Q(A)$ at the 200th trial

Without block

\[ 0 1 2 3 4 5 6 7 8 \]
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 \( \gamma \), 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 \( R_{i} \) was set to be a constant \( R_{i} = 1 \), and eight cases with different amounts of delayed reward \( R_{d} = 1, 2, \ldots, 8 \) were simulated (200 trials \( \times \) 500 simulations for each case). (C-G) Simulation results. (C) Percentage of choosing delayed reward \( R_{d} \) 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 \( Q(A_{1}) \) \( \pm \) 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 case of RPE and negative RPE, respectively, and they are thus expected to positively or negative 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 \( (A_{1} \text{ or } A_{2}) \), 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.

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 \( S_{3} \) 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-
environment. 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., 2013). 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), corticostriatal (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., 2016 K. Morita and Y. Kawaguchi
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.

As shown in the Results, the CS-TD model predicts that dMSNs and iMSNs 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-recombine 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 CPu/PT cells encoding newer and older information, respectively. In particular, CCS and CPu/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 CPu/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 CPu/PT cells. In that case, the temporal difference in the encoded information between CCS and CPu/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 CPu/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 (regardless of the mechanism of RPE calculation). (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. a 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; CPu/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-pallidal 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) Intrathecopetal corticostriatal neurons equally excite striatonic 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. Nature, 471, 207–213.

Brock, R., Shin, J.H., Kaplan, A.R., Dobi, A., Markey, E., Kramer, P.F., Gre- nel, C.M., Christensen, C.H., Adrover, M.F. & Alvarez, V.A. (2013) Strengthening the accumbal indirect pathway promotes resilience to compulsive cocaine use. Nature, 506, 677–681.

Collins, A.G. & Frank, M.J. (2014) Opponent actor learning (OpAL): model- ing feature-based representation of state in the basal ganglia. J. Neurosci., 34, 8341–8350.

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.

Hirai, Y., Morishima, M., Karube, F. & Kawaguchi, Y. (2012) Specialized subthalamic nucleus-neurotransmitter function: adaptive gain and optimal performance. Neuron, 74, 351–362.

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

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. J. Neurosci., 35, 1397–1406.
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., Ohyma, T., Kreider, J.C., Rieusech, F. & Maud, M.D. (2009) Interactions between prefrontal cortex and cerebellum revealed by trace eyelid conditioning. Learn. Memory, 16, 86–95.

Keeler, J.F., Pretzoll, 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., Suh, 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., Shepherd, N.D., Elliott, H.L., Xie, T., Gerfen, C.R. & Sabatini, B.L. (2015) A direct pathway and dendral striatum in instrumental reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. J. Neurosci., 35, 457–467.

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.

Kress, G.J., Yamawaki, N., Wokosin, D.L., Wickersham, I.R., Shepherd, N.D., Elliott, H.L., Xie, T., Gerfen, C.R. & Sabatini, B.L. (2015) A direct pathway and dendral striatum in instrumental reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. J. Neurosci., 35, 457–467.

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 differentiation and projection-specific corticostriatal 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.

Morris, G., Nevet, A., Arkadis, D., Vadaia, E. & Bergman, H. (2006) Mid-brain dopamine neurons encode decisions for future action. Nat. Neurosci., 9, 1059–1063.

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.

Nambar, A., Tokuno, H. & Takada, M. (2002) Functional significance of the corticostriatal-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. (1999) 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.L. & 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.

Roosch, 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. & Kiumura, 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., Berezovski, V.K., Johnson, C.A., Kingery, N.D., Elliott, H.L., Xie, T., Gerfen, C.R. & Sabatini, B.L. (2015) A direct dopamineergic output from the basal ganglia to frontal cortex. Nature, 521, 85–89.

Schroll, H., Vityat, J. & Hanker, 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.
