Systems/Circuits

Dopaminergic Control of Motivation and Reinforcement Learning: A Closed-Circuit Account for Reward-Oriented Behavior

Kenji Morita,[1,2] **Mieko Morishima,**[3,4,5] **Katsuyuki Sakai,**[2] **and Yasuo Kawaguchi**[3,4,5]

[1]Physical and Health Education, Graduate School of Education and [2]Department of Cognitive Neuroscience, Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan, [3]Division of Cerebral Circuitry, National Institute for Physiological Sciences, Okazaki 444-8787, Japan, [4]Department of Physiological Sciences, Graduate University for Advanced Studies, Okazaki 444-8787, Japan, and [5]Japan Science and Technology Agency, Core Research for Evolutional Science and Technology, Tokyo 102-0076, Japan

Humans and animals take actions quickly when they expect that the actions lead to reward, reflecting their motivation. Injection of dopamine receptor antagonists into the striatum has been shown to slow such reward-seeking behavior, suggesting that dopamine is involved in the control of motivational processes. Meanwhile, neurophysiological studies have revealed that phasic response of dopamine neurons appears to represent reward prediction error, indicating that dopamine plays central roles in reinforcement learning. However, previous attempts to elucidate the mechanisms of these dopaminergic controls have not fully explained how the motivational and learning aspects are related and whether they can be understood by the way the activity of dopamine neurons itself is controlled by their upstream circuitries. To address this issue, we constructed a closed-circuit model of the corticobasal ganglia system based on recent findings regarding intracortical and corticostriatal circuit architectures. Simulations show that the model could reproduce the observed distinct motivational effects of D_1- and D_2-type dopamine receptor antagonists. Simultaneously, our model successfully explains the dopaminergic representation of reward prediction error as observed in behaving animals during learning tasks and could also explain distinct choice biases induced by optogenetic stimulation of the D_1 and D_2 receptor-expressing striatal neurons. These results indicate that the suggested roles of dopamine in motivational control and reinforcement learning can be understood in a unified manner through a notion that the indirect pathway of the basal ganglia represents the value of states/actions at a previous time point, an empirically driven key assumption of our model.

Introduction

Dopamine has been suggested to control motivation and reward-seeking behavior (Robbins and Everitt, 1996; Berridge and Robinson, 1998; Dayan and Balleine, 2002; McClure et al., 2003; Niv, 2007). As a direct evidence, application of dopamine receptor antagonists in the striatum has been shown to slow the subject's behavior (Salamone and Correa, 2002; Nakamura and Hikosaka, 2006), with distinct effects observed for the antagonists of D1- and D2-type dopamine receptors (D1Rs and D2Rs), which are expressed in distinct populations of striatal medium spiny neurons (MSNs) projecting to the "direct" (dMSNs) and "indirect" (iMSNs) pathways of the basal ganglia, respectively (Gerfen and Surmeier, 2011). Clarifying mechanisms of such pharmacological effects is likely to help elucidating the exact roles of dopamine in motivational control, and neural circuit modeling has been used as one of the powerful approaches (Frank et al., 2004; Hong and Hikosaka, 2011). However, these models are still not self-contained in a sense that responses of either the dopamine neurons or their hypothesized upstream globus pallidus (GP) neurons were presumed rather than explained by inputs from the rest part of the circuit.

Along with the suggested roles in motivational control, dopamine has also been suggested to be centrally involved in reinforcement learning. Specifically, neurophysiological studies with computational perspectives have revealed that phasic response of dopamine neurons appears to represent the temporal-difference (TD) reward prediction error (Montague et al., 1996; Schultz et al., 1997; Bayer and Glimcher, 2005; Roesch et al., 2007), a key quantity defined in reinforcement learning algorithms (Sutton and Barto, 1998). Together with the finding that corticostriatal synapses are plastically modified according to phasic dopamine response (Reynolds et al., 2001), the central role of dopamine in reinforcement leaning has become widely accepted (Glimcher,

Received Sept. 28, 2012; revised March 21, 2013; accepted March 25, 2013.

Author contributions: K.M. designed research; K.M. performed research; K.M., M.M., K.S., and Y.K. analyzed data; K.M., K.S., and Y.K. wrote the paper.

This work was supported by Ministry of Education, Science, Sports, and Culture of Japan Grant-in-Aid for Scientific Research on Innovative Areas "Mesoscopic Neurocircuitry" 23115505 (K.M.), Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Young Scientists(B) 24700312 (K.M.) and 24700338 (M.M.) and Grant-in-Aid for Scientific Research 21240030 (Y.K.), JSPS Funding Program for Next Generation World-Leading Researchers Grant LS030 (K.S.), and Japan Science and Technology Agency, Core Research for Evolutional Science and Technology (Y.K.). We thank the anonymous reviewers for their helpful comments on this manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Kenji Morita, Physical and Health Education, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: morita@p.u-tokyo.ac.jp.

DOI:10.1523/JNEUROSCI.4614-12.2013

 0270-6474/13/338866-25$15.00/0

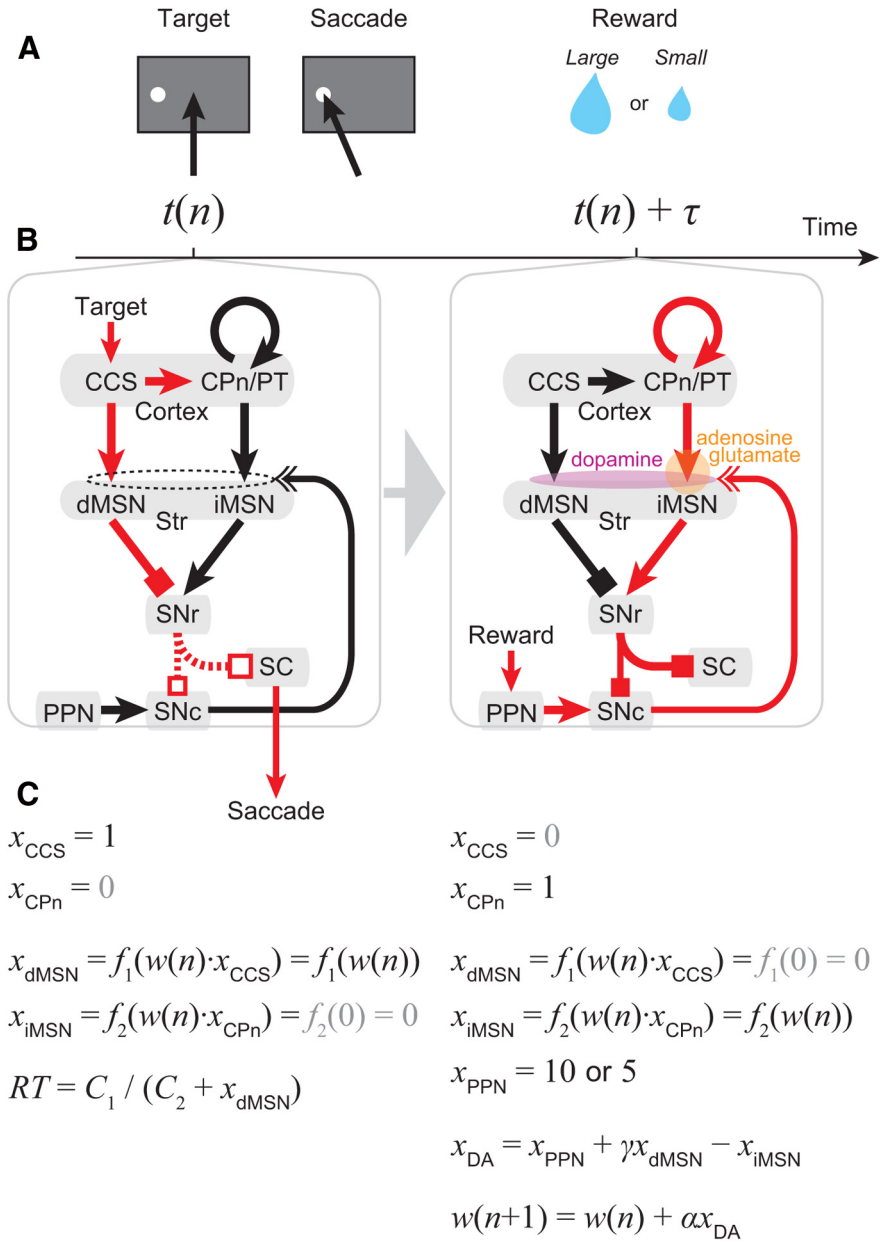

Figure 1. Simulated visually guided saccade task and hypothesized corticobasal ganglia circuit. ***A***, The simulated task. On each trial, a target appears pseudorandomly at the left or right, and a correct response is followed by either large or small reward. Contingency between the target location and the reward amount is fixed for each block (20–28 trials) and alternated across blocks. ***B***, Hypothesized structure and operation of the corticobasal ganglia circuit. Target location is represented by a subset of a major subtype of corticostriatal neurons, CCS cells, that predominantly target striatonigral (direct pathway) dMSNs (Reiner et al., 2010) (left). These CCS cells activate dMSNs, which then inhibit the SNr and thereby disinhibit the SC to initiate a saccade. Meanwhile, the CCS cells also activate, via unidirectional projections (Morishima and Kawaguchi, 2006), another major subtype of corticostriatal neurons, CPn/PT cells, which predominantly target striatopallidal (indirect pathway) iMSNs (Reiner et al., 2010). These CPn/PT cells sustain their activity via strong facilitatory recurrent excitation (Morishima et al., 2011), whereas the activity of CCS cells declines because of their weaker and depressive recurrent excitation. When reward is obtained (right), dopamine neurons in the SNc receive excitatory inputs representing the value (amount) of the obtained reward from the neurons in the PPN. The SNc dopamine neurons also receive inhibitory inputs from the collaterals of SNr neurons, which are disinhibited by the iMSNs downstream of the still active CPn/PT cells. Given that the CPn/PT cells represent the location of the saccadic target and that the iMSNs represent the reward value predicted from that location, the SNr → SNc inhibition represents the reversed-sign predicted reward value. Thus, together with the input from the PPN, the dopamine neurons compute reward prediction error. The resulting phasic dopamine response induces proportional plastic changes of the synaptic strength between the CCS cells and dMSNs so that the activity of dMSNs can represent updated reward value predictions. At the synapses between the CPn/PT cells and iMSNs, sustained inputs from the CPn/PT cells cause adenosine accumulation and glutamate spillover. Phasic increase in dopamine then stimulates the signaling cascade downstream of the adenosine A_{2A} receptors, leading to LTP, whereas phasic decrease in dopamine causes LTD through the signaling cascade downstream of mGluR5. In consequence, the activity of iMSNs can also represent updated reward value predictions. Notably, molecular mechanisms of plasticity induction, as well as structures between the striatum (Str) and the dopamine neurons, including the SNr, were not explicitly modeled, and our model can be also compatible with recent findings that the dopamine neurons are regulated by the LHb neurons, which are then driven by GPb (Fig. 7*A*,*B*). Also notably, although not explicitly illustrated here, we implicitly assumed that the striatum and the SNr are topographically organized and only a portion of neurons having the corresponding saccadic response fields, rather than an entire population, is activated in the task. ***C***, Model equations used for the simulations shown in Figures 2, 4, and 5. x_{CCS} and x_{CPn} indicate the activity of the CCS cells and the CPn/PT cells that represent the left-target location (value of 1 is arbitrary; for details, see Materials and Methods), x_{dMSN} and x_{iMSN} indicate the population activity of dMSNs and iMSNs, respectively, x_{PPN} indicates the activity of the PPN neurons, and x_{DA} indicates the response of dopamine neurons compared with its baseline activity. $w(n)$ indicates the strength of the connections between the CCS cells and dMSNs and those between the CPn/PT cells and iMSNs in the nth left-target trial [note that we assumed that CCS–dMSN and CPn/PT–iMSN connections are updated in the same way and so described their strength by the single variable $w(n)$]. f_1 and f_2 represent the input–output functions of dMSNs and iMSNs, respectively: these functions are assumed to be a threshold–linear function under the presence of tonic dopamine (Fig. 3). α represents the learning rate. γ represents the (relative) strength/efficacy of the direct pathway over the indirect pathway, and it corresponds to the time discount factor defined in the TD reinforcement learning algorithms (Fig. 6*C*). The gray parts in the equations will be modified in the elaborated model used for the simulations shown in Figure 6.

2011), leading to a proposal of a closed neural circuit model (Potjans et al., 2011). Despite these progressions, however, exact circuit mechanisms for the computation of TD error in the upstream of dopamine neurons had remained quite elusive, making it difficult to combine such a circuit model of reinforcement learning with the models of motivational control with distinct D_1R and D_2R effects as introduced above.

Recently, detailed anatomical and physiological features of the corticostriatal circuit have become revealed. Specifically, it has been shown that dMSNs and iMSNs are predominantly targeted by distinct corticostriatal neurons, named crossed-corticostriatal (CCS) cells and corticopontine/pyramidal-tract (CPn/PT) cells, respectively (Lei et al., 2004; Reiner et al., 2010). Moreover, it has been demonstrated that the CCS cells unidirectionally project to the CPn/PT cells (Morishima and Kawaguchi, 2006), and the CPn/PT cells (but not the CCS cells) possess strong facilitatory recurrent excitation (Morishima et al., 2011). These features have led us to conjecture how the activity of dopamine neurons is regulated and can represent TD reward prediction error (Morita et al., 2012). In the present study, we constructed a closed neural circuit model based on this conjecture and

tried to simultaneously explain the observed motivational effects of dopamine receptor antagonists and the suggested roles of dopamine in reinforcement learning by simulating two behavioral tasks.

Materials and Methods

Simulated saccade task. We first simulated a visually guided saccade task (Fig. 1*A*) used in an experimental study (Nakamura and Hikosaka, 2006). In the experiment, a visual stimulus (saccadic target) appeared at the left or the right of the screen on each trial, and the subject (monkey) was required to make a saccade toward the target. If the subject made a correct response, a liquid reward was given after 100 ms of fixation at the target. There were two kinds of reward amount, "large" (0.4 ml) or "small" (0.05 or 0 ml), each of which was associated with either left or right target; contingency between the target location and the reward amount was fixed for an individual task block consisting of 20–28 trials, and "left-large" and "left-small" blocks were switched alternately without a preceding cue for a switch. The target location (left and right) was pseudorandomly determined in each trial. We incorporated these features into our simulations, although there are several points that differ from the experimental study as we describe below. We constructed two models: (1) a simpler one, which modeled neuronal activity only at the timings of target presentation and reward reception, and (2) an elaborated one, which modeled intertrial intervals as well. In both of the models, we did not directly model the amount of reward or neural processes for reward sensation/consumption but instead assumed two different levels of reward-representing inputs to the dopamine neurons for the large- and small-reward conditions (see below). Also, we did not model the two different amounts of reward for the small-reward condition (0.05 or 0 ml); in the experiment (Nakamura and Hikosaka, 2006), trials with 0.05 ml and those with 0 ml were not analyzed separately. With the simpler model, we simulated only the trials with a left target for simplicity and in the same manner as the previous modeling study (Hong and Hikosaka, 2011). With the elaborated model, we simulated both left-target and right-target trials, although only the results for the left-target trials were shown in the figures unless otherwise mentioned. In the following, we first describe the details of the simpler model and, thereafter, those about the elaborated one.

Simulated neural circuit for reward-oriented saccade. A series of experimental studies have revealed that the cortex, basal ganglia, and the dopamine neurons in the substantia nigra pars compacta (SNc) play essential roles in learning and execution of reward-associated saccade tasks (Kawagoe et al., 2004; Takikawa et al., 2004; Hikosaka et al., 2006; Nakamura and Hikosaka, 2006). We constructed a computational model of the corticobasal ganglia–SNc circuit according to the conjecture (Morita et al., 2012) derived from recent anatomical and physiological findings (Fig. 1*B*). There are two distinct types of corticostriatal neurons, the CCS cells and the CPn/PT cells, and there exist unidirectional projections from the CCS cells to the CPn/PT cells (Morishima and Kawaguchi, 2006) and strong facilitatory recurrent excitation only among the CPn/PT cells (Morishima et al., 2011). The CCS cells and the CPn/PT cells predominantly project to the striatal dMSNs and iMSNs, respectively (Lei et al., 2004; Reiner et al., 2010), which presumably upregulate and downregulate the dopamine neurons in the SNc (Aggarwal et al., 2012; Morita et al., 2012) via the substantia nigra pars reticulata (SNr) (Tepper and Lee, 2007). The SNc dopamine neurons receive significant inputs also from other structures (Watabe-Uchida et al., 2012), including excitation from the pedunculopontine nucleus (PPN) (Mena-Segovia et al., 2004) and inhibition from the striatal striosomes (Gerfen et al., 1985; Paladini et al., 1999; Fujiyama et al., 2011; Watabe-Uchida et al., 2012). The former input could convey information of actually obtained reward, because it has been shown (Okada et al., 2009) that a population of neurons in the PPN represents such information. We did not incorporate the latter, striosomal input to the dopamine neurons or many other known connections. Exploring how these unincorporated connections can be related to the circuit that we modeled is an important future issue.

We assumed that phasic dopamine response induces proportional changes of the strengths of connections between CCS cells and dMSNs and those between CPn/PT cells and iMSNs via plasticity mechanisms; in the case of synapses on the iMSNs, presumably together with adenosine acting on the A_{2A} receptors and/or glutamate acting on the metabotropic glutamate receptor 5 (mGluR5) that are assumed to be synaptically accumulated during sustained inputs from CPn/PT cells (see below). Phasic dopamine response may also gradually change the level of tonic dopamine (cf. Niv et al., 2007). However, in the saccade task that we simulated (Nakamura and Hikosaka, 2006), as well as the other task that we also simulated as we describe below (Roesch et al., 2007; Takahashi et al., 2011), such changes across blocks would not be very significant, because large- and small-reward trials were intermingled and also because the frequency of phasic release (on reward reception) is relatively low (each task trial takes >4 s) and seems not to vary much compared with self-paced free-operant tasks. We thus did not consider changes in tonic dopamine; nevertheless, we did consider the effects of tonic dopamine on the responsiveness of dMSNs/iMSNs and their blockade by antagonists as we explain below.

Because the experimental results (Kawagoe et al., 2004; Hikosaka et al., 2006; Nakamura and Hikosaka, 2006) suggest that the caudate is especially involved in the learning and execution of reward-associated saccade tasks, "striatum" in our models for the saccade task basically refers to the caudate. Likewise, "cortex" refers to the areas that project to the caudate and exhibit saccade-related activity, e.g., the frontal eye field (FEF), and we assumed that there exist CCS cells and CPn/PT cells in those areas. However, other striatal regions could potentially operate in similar ways given the suggested common architecture across the entire striatum (Pennartz et al., 2011). There have been various types of neurons in the striatum (Hikosaka et al., 2006), and not all of them are incorporated into our model; in particular, neurons showing block-indicating activity before the presentation of the target (Watanabe and Hikosaka, 2005; Hikosaka et al., 2006) were not incorporated.

We assumed the following set of equations (Fig. 1*C*) describing the time (t)-dependent neuronal activity of cortical CCS cells and CPn/PT cells that represent the left-target location [$x_{CCS}(t)$ and $x_{CPn}(t)$], the activity of striatal dMSNs and iMSNs [$x_{dMSN}(t)$ and $x_{iMSN}(t)$], the activity of the PPN neurons that represent obtained reward [$x_{PPN}(t)$], the response of SNc dopamine neurons compared with its baseline activity [$x_{DA}(t)$], and the strength of connections between the CCS cells and the dMSNs and those between the CPn/PT cells and the iMSNs [$w(n)$, where n represents the index of left-target trials, i.e., $w(n)$ represents the strength at the nth left-target trial]. Notably, $w(n)$ represents both the CCS–dMSNs and CPn/PT–iMSNs connection strengths, because we assumed that the synaptic strength between CCS cells and dMSNs and the strength between CPn/PT cells and iMSNs are modified in effect similarly (see below). At the timing of target presentation in the nth left-target trial [$t = t(n)$] in a given block,

$$x_{CCS}(t(n)) = 1.$$

The CCS cells that represent the left-target location become active in response to a presentation of the left target. Notably, it is arbitrary to set this activity to what value; we set this to 1 for simplicity, but we can equivalently set this to 10, for example, and scale down $w(n)$ and also α (learning rate; see below) to 1/10 of the original values:

$$x_{CPn}(t(n)) = 0.$$

The CPn/PT cells that represent the left-target location presumably do not show immediate response to a presentation of the target but rather gradually become activated by the CCS cells via the unidirectional connections (Morishima and Kawaguchi, 2006), and so we assumed that their activity is 0 at time $t(n)$. More precisely, however, other CPn/PT cells could represent a state preceding the target presentation and drive iMSNs. This point was not incorporated into the model considered here for simplicity but was taken into account in our elaborated model described below:

$$x_{dMSN}(t(n)) = f_1(w(n) \times x_{CCS}(t(n))) = f_1(w(n) \times 1) = f_1(w(n)).$$

$w(n) \times x_{CCS}(t(n))$ $(= w(n))$ represents the inputs from the CCS cells to the dMSNs. f_1 is the input–output function of the dMSNs under the presence of tonic dopamine (see below), and $x_{dMSN}(t(n))$ presumably represents the reward value (reward amount) predicted from the left-

target location under the condition in which f_1 operates in the suprathreshold linear regimen and is not affected by D_1 antagonist:

$$x_{\mathrm{iMSN}}(t(n)) = f_2(w(n) \times x_{\mathrm{CPn}}(t(n))) = f_2(w(n) \times 0) = f_2(0) = 0.$$

f_2 is the input–output function of the iMSNs under the presence of tonic dopamine (see below), which is assumed to take 0 when input is 0:

$$x_{\mathrm{PPN}}(t(n)) = 0.$$

This indicates that the visual stimulus (saccadic target) itself is neutral (not rewarding):

$$x_{\mathrm{DA}}(t(n)) = x_{\mathrm{PPN}}(t(n)) + \gamma x_{\mathrm{dMSN}}(t(n)) - x_{\mathrm{iMSN}}(t(n)).$$

This relationship was derived from the hypothesized upstream circuitry of dopamine neurons (Fig. 1*B*) (Morita et al., 2012). γ represents the (relative) strength/efficacy of the direct pathway over the indirect pathway (its value does not affect the results presented in Figs. 2, 4, 5). Notably, it has been suggested that narrower dynamic range of negative dopamine response than positive response can significantly affect learning (Potjans et al., 2011), but such asymmetry was not considered in the present study; considering it in our model could be a good theme for future research.

At the timing of reward reception in the nth left-target trial [$t = t(n) + \tau$, in which τ includes a reaction time, period for fixation at target (100 ms was required in the experiment) (Nakamura and Hikosaka, 2006) and time for liquid reward delivery and sensation/consumption],

$$x_{\mathrm{CCS}}(t(n) + \tau) = 0.$$

The target-induced activity of the CCS cells is assumed to decline, because the recurrent excitation among CCS cells is relatively weak and entails short-term synaptic depression (Morishima et al., 2011). Therefore, we assumed $x_{\mathrm{CCS}}(t(n) + \tau) = 0$. More precisely, however, other CCS cells could represent a new state at time $t(n) + \tau$ and drive dMSNs. This point was not incorporated into the model considered here for simplicity but was taken into account in our elaborated model described below:

$$x_{\mathrm{CPn}}(t(n) + \tau) = 1.$$

We assumed this because the CPn/PT cells presumably become active by the input from the CCS cells via the unidirectional connections (Morishima and Kawaguchi, 2006) and then sustain activity via strong recurrent excitation (Morishima et al., 2011):

$$x_{\mathrm{dMSN}}(t(n) + \tau) = f_1(w(n) \times x_{\mathrm{CCS}}(t(n) + \tau)) = f_1(w(n) \times 0) = f_1(0) = 0.$$

f_1 (as well as f_2) is assumed to take 0 when input is 0 (see below).

$$x_{\mathrm{iMSN}}(t(n) + \tau) = f_2(w(n) \times x_{\mathrm{CPn}}(t(n) + \tau)) = f_2(w(n) \times 1) = f_2(w(n)).$$

This is equal to $x_{\mathrm{dMSN}}(t(n))$, thus representing the reward value (reward amount) predicted from the left-target location, under the condition in which f_1 and f_2 are in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists (see below):

$$x_{\mathrm{PPN}}(t(n) + \tau) = 5 \text{ or } 10.$$

The 5 and 10 values correspond to the small- and large-reward conditions, respectively:

$$x_{\mathrm{DA}}(t(n) + \tau) = x_{\mathrm{PPN}}(t(n) + \tau) + \gamma x_{\mathrm{dMSN}}(t(n) + \tau) - x_{\mathrm{iMSN}}(t(n) + \tau).$$

This relationship (Morita et al., 2012) indicates that the dopaminergic neuronal response represents reward prediction error or, more specifically, TD error defined in the TD learning (Sutton and Barto, 1998), and the parameter γ, representing the (relative) strength/efficacy of the direct pathway over the indirect pathway, corresponds to the degree of time discount for future rewards, namely, time discount factor.

In reference to empirical findings (Kawagoe et al., 2004) and the results of the previous modeling study (Hong and Hikosaka, 2011), we assumed a simple quasi-inverse relationship between the saccadic reaction time [$RT(n)$] and the stimulus-induced response of the dMSNs:

$$RT(n) = C_1/(C_2 + x_{\mathrm{dMSN}}(t(n))),$$

where C_1 and C_2 are constants and were set to 3000 and 6, respectively, in the simulations shown in Figures 2, 4, and 5.

Notably, as shown in the above, we did not explicitly model structures between the striatum and the dopamine neurons, namely, the SNr, the external segment of the GP (GPe), and the subthalamic nucleus. For one thing, this was for simplicity; however, with such abstraction, our model could be also compatible with recent findings (Matsumoto and Hikosaka, 2007; Hong and Hikosaka, 2008; Hong et al., 2011; Shabel et al., 2012) that the dopamine neurons, in the SNc and the ventral tegmental area (VTA), are regulated by the lateral habenula (LHb) neurons, which are then driven by the border region of GP (GPb). Specifically, we assumed that the same corticostriatal circuit mechanism for computing the TD of values of the current and previous states/actions (Morita et al., 2012) can operate with two different circuits connecting to the dopamine neurons: (1) one via the SNr (see Fig. 7*A*) and (2) the other via the GPb and the LHb (see Fig. 7*B*) (see Results, Reward prediction error computation in parallel with action selection and/or execution).

Presumed effects of dopamine receptor antagonists. Existence of tonic dopamine is considered to affect the relationship between the strength of cortical inputs and the firing rate of MSNs, in different ways for dMSNs and iMSNs, and application of D_1- or D_2-type dopamine receptor antagonist presumably "demodulates" such a tonic dopamine-modulated relationship to a certain extent. For simplicity, we assumed the following threshold–linear function for the input–output functions of dMSNs and iMSNs under the condition with a presumed certain level of tonic dopamine and without antagonists (see Fig. 3*A*,*B*, black lines):

$$f_1(I) = 0 \ (I \leq \theta),$$

$$I - \theta \ (\theta < I),$$

$$f_2 \ (I) = 0 (I \leq \theta),$$

$$I - \theta \ (\theta < I),$$

where I is input to MSNs, and θ represents a threshold level, which was set to 5 in the simulations. The actual relationship between the input strength and the output firing rate under this condition may entail a certain degree of saturating nonlinearity, but we assumed the above threshold–linear function for simplicity.

Dopamine is known to upregulate the activity of dMSNs via the activation of D_1Rs and downregulate the activity of iMSNs via the activation of D_2Rs (Gerfen and Surmeier, 2011), although exact ways of modulation *in vivo* remain elusive. As for the D_1Rs, it has been suggested that their activation enhances the NMDA current more significantly than the AMPA current (Levine et al., 1996; Flores-Hernández et al., 2002; Moyer et al., 2007). Because the NMDA receptor has voltage dependence (Schiller and Schiller, 2001), the enhancement of the NMDA current would become more prominent when dMSNs receive stronger glutamatergic (cortical) inputs. Therefore, we assumed that the response of dMSNs to strong (but not weak) inputs, which would normally be enhanced by tonic dopamine, is attenuated by D_1 antagonist (see Fig. 3*A*). In the presence of D_1 antagonist (see Fig. 3*A*, gray),

$$f_1(I) = 0 \ (I \leq 5),$$

$$I - 5 \ (5 < I \leq 12),$$

$$7 + 0.6 \ (I - 12)(12 < I).$$

Conversely, regarding the D_2Rs, it was originally suggested that their activation suppresses the AMPA current more significantly than the NMDA current (Levine et al., 1996; Flores-Hernández et al., 2002; Hernández-Echeagaray et al., 2004; Moyer et al., 2007), but later studies (Azdad et al., 2009; Higley and Sabatini, 2010) have shown that D_2R activation does reduce the NMDA receptor-mediated excitation [and in fact, Higley and Sabatini (2010) has shown that D_2R activation did not affect non-NMDA synaptic currents, apparently contradictory to the results mentioned above]. At the same time, however, it was also shown that, if A_{2A} adenosine receptors, which are predominantly expressed in iMSNs, are simultaneously activated, such a suppression of the NMDA receptor-mediated excitation by D_2R activation can be fully blocked (Azdad et al., 2009; Higley and Sabatini, 2010). It has been suggested (Cunha, 2001; Schiffmann et al., 2007) that the formation of extracellular adenosine results from the action of ecto-nucleotidases on ATP released with neurotransmitters, including glutamate, and thus the synaptic pool of adenosine reflects neuronal firing. Together, it is conceivable that the activation of D_2Rs reduces the AMPA and/or NMDA receptor-mediated excitation when iMSNs receive weak cortical inputs and there exist a moderate amount of adenosine, whereas D_2R activation could still reduce the AMPA current but would never reduce the NMDA receptor-mediated excitation when iMSNs receive strong cortical inputs and there exist a high amount of adenosine. Based on these considerations, we assumed that the response of iMSNs to weak (but not strong) inputs, which would normally be suppressed by tonic dopamine, is enhanced by D_2 antagonist (see Fig. 3*B*). In the presence of D_2 antagonist (see Fig. 3*B*, gray),

$$f_2(I) = 0 \ (I \leq 2),$$

$$7 + 0.7 \ (I - 12) (2 < I \leq 12),$$

$$I - 5 \ (12 < I).$$

Notably, in our simulations shown in Figures 2, 4, and 5, $I = w(n)$ [i.e., input to dMSNs at time $t(n)$ and input to iMSNs at time $t(n) + \tau$] was always, except for the initial transient, near the central part of the input range shown in Figure 3, *A* and *B*, and therefore both D_1 and D_2 antagonists were able to properly cause their effects. Changing the assumptions about the effects of the antagonists on the input–output functions, or the magnitude of rewards, can significantly alter the main results of this study. Nevertheless, given the observed cortical neural representation of relative (rather than absolute) preference of rewards (Tremblay and Schultz, 1999), matching between the range of reward amounts relevant in a given situation (task) and the dynamic range of value-representing neurons in the brain, including the MSNs, could actually be achieved via certain mechanisms.

Presumed dopamine-dependent corticostriatal plasticity. We assumed that the reward value (reward amount) predicted from the left-target location is stored in the strength of corticostriatal synapses, in the manner as described above, and it is updated through plasticity mechanisms depending on phasic dopamine response after reward reception as in the following equation:

$$w(n + 1) = w(n) + \alpha x_{DA}(t(n) + \tau).$$

The parameter α represents the learning rate, and it was set to 0.75, which is close to an empirically estimated value (0.7) in a study using a different task (Bayer and Glimcher, 2005). We checked that the central features of our simulation results shown in Figures 2, 4, and 5 were successfully reproduced even if the learning rate was set to 0.6 or 0.9 (data not shown). We assumed the same form of dopamine-dependent changes of corticostriatal inputs as above for both the dMSNs and iMSNs. It might appear odd to make such an assumption, given the distinct properties of the D_1Rs and D_2Rs that are separately expressed in dMSNs and iMSNs (Gerfen and Surmeier, 2011). However, according to our hypothesis (Fig. 1*B*), the synapses on dMSNs and those on iMSNs are in fact under quite different conditions at time $t(n) + \tau$. Specifically, whereas the synapses on iMSNs continuously receive inputs from the CPn/PT cells that sustain activity, those on dMSNs no longer receive inputs from the CCS cells representing the same target stimulus because those CCS cells presumably have already become inactive. It has been suggested (Shen et al., 2008) that, different from the case of dMSNs, induction of long-term potentiation (LTP) or long-term depression (LTD) in iMSNs requires activation of the A_{2A} adenosine receptors or mGluR5, respectively. These two receptor types are suggested to form oligomers with D_2Rs (Cabello et al., 2009), and they are presumably activated by adenosine, which is generated around synapses from ATP released with glutamate (Cunha, 2001; Schiffmann et al., 2007) and glutamate that is spilled over from synapses (Mitrano et al., 2010), respectively. We conjectured that the sustained inputs from CPn/PT cells cause adenosine accumulation and glutamate spillover at the synapses on iMSNs, and such adenosine and glutamate can activate the A_{2A} receptors and mGluR5, respectively (Fig. 1*B*, right, orange circle). Moreover, because it has been suggested that effective generation of a response downstream of the A_{2A} receptors needs a stimulation of D_2Rs (Azdad et al., 2009), we conjectured that phasic increase in dopamine stimulates the A_{2A} receptor-signaling cascade leading to LTP, whereas phasic decrease in dopamine results in LTD presumably through the mGluR5-signaling cascade. Based on these conjectures, we assumed that the above equation effectively holds for both the synapses on dMSNs and those on iMSNs. Notably, however, it has been shown that, at least under a certain condition, A_{2A} receptor agonist induced LTP in the presence of D_2 antagonist (Shen et al., 2008, their Fig. 1*F*). Possibly, afferent stimulation applied in that study caused phasic dopamine release, and its effect on plasticity was not completely blocked by D_2 antagonist (although the responsiveness of iMSNs was likely to be significantly affected by D_2 antagonist). However, exploring how the assumption of our model can be well reconciled with these experimental results remains as an important future issue.

Plasticity induction depending on phasic dopamine response has been demonstrated *in vivo* (Reynolds et al., 2001). However, the synapses stimulated by contralateral cortical stimulation in that study would be mainly between CCS cells and dMSNs, because CCS cells but not CPn/PT cells target contralateral striatum (Reiner et al., 2010). For the synapses on dMSNs, a precise intracellular mechanism has also been proposed (Nakano et al., 2010). However, whether phasic decrease in dopamine induces LTD as assumed in the above was not tested previously by Reynolds et al. (2001) or Shen et al. (2008); in the latter study, effect of D_1 antagonist was examined, but it can be different from effects of phasic dopamine decrease. On the other hand, involvement of phasic dopamine response in plasticity induction in the synapses on iMSNs remains unclear. The abovementioned study by Shen et al. (2008) has shown that tonic activation of D_2Rs (by bath application of D_2 agonist *in vitro*) is unnecessary for LTP induction and would rather induce LTD in those synapses. However, again, afferent stimulation applied in that study possibly caused phasic dopamine release, and its effect was possibly masked by bath application of D_2 agonist [possible occurrence of such a masking *in vivo* has been discussed previously (Klein et al., 2007)]. Potential difference between bacterial artificial chromosome transgenic mice and wild-type animals (Bagetta et al., 2011; Kramer et al., 2011; Chan et al., 2012; Nelson et al., 2012) may also need to be carefully considered.

Overall, our assumption on plasticity could be in line with, or at least not inconsistent with, several experimental results as we explained above, but there remain divergences and limitations that need to be clarified in the future. An important limitation of our model, which would especially affect plasticity, is that it modeled neural activity as a continuous variable and did not model individual spike timings. In fact, a major finding by the *in vitro* study mentioned above for plasticity (Shen et al., 2008) was that the direction of induced plasticity (i.e., LTP or LTD) critically depends on the precise timings of presynaptic and postsynaptic neuronal firings (as known as spike-timing-dependent plasticity). Our model, lacking spikes, cannot be directly compared with their experiments. It is expected that more elaborated models using spiking neuron models will be developed in the future, e.g., combining the existing spiking neuron model for reinforcement learning (Potjans et al., 2011) with the circuit architecture that we proposed in this study. Also notably, not only the temporal resolution but also the spatial resolution of the model would need to be improved, given recent suggestions that plasticity might actually be regulated not by somato-axonic firings but instead by dendritic

local spikes (Poirazi and Mel, 2001; Golding et al., 2002; Morita, 2009; Legenstein and Maass, 2011) in pyramidal cells and also possibly in the MSNs as the author of the abovementioned plasticity study himself recently pointed out (Surmeier et al., 2011).

Numerical simulations. We set the initial value of $w(n)$ to 0 and simulated successive 501 task blocks in three conditions: (1) without dopamine receptor antagonists; (2) with D_1 antagonist; and (3) with D_2 antagonist. The simulations were conducted by using MATLAB (MathWorks), and the presented data are the averages of the last 500 blocks for each condition (the initial block was not included because there was initial transient). For the simulations, we wrote codes (driven equations) for 1/10 scaled-down values of the variables (except for x_{CCS} and x_{CPn}), and then the results were 10 times scaled up when they were illustrated (this is just for practical reasons, and it should be mathematically equivalent to writing codes for the variables without scaling).

Elaborated model for the saccade task. In the simple model described so far, we assumed that the activity of dMSNs at the timing of reward, which represents the predicted value of the "end-of-trial" state, was 0. However, given an empirical suggestion (Enomoto et al., 2011) that the response of dopamine neurons can reflect multiple rewards over multiple trials, such an assumption may not be very appropriate. Therefore, to examine effects of reward prediction (expectation) over multiple trials, we constructed an elaborated model for the same saccade task (Nakamura and Hikosaka, 2006) (see Fig. 6*A–C*). At each time step (t_i), the subject is assumed to be in one of the states shown in Figure 6*A*. A subset of CCS cells is assumed to represent that state, $S(t_i)$ (see Fig. 6*B*), and the input from these CCS cells to dMSNs, denoted by $I(S(t_i))$, is assumed to represent the predicted value of state $S(t_i)$ [denoted by $V(S(t_i))$], shifted by the amount of the threshold of the input–output function, i.e., $V(S(t_i)) = I(S(t_i)) - \theta$. The activity of the dMSNs is assumed to be determined by imposing the input–output function f_1 (the same function defined above) on $I(S(t_i))$, i.e., $f_1(I(S(t_i)))$, which is equal to $V(S(t_i))$ under the condition in which f_1 operates in the suprathreshold linear regimen and is not affected by D_1 antagonist. Meanwhile, a subset of CPn/PT cells is assumed to represent the state of the subject at the previous time step, i.e., $S(t_{i-1})$ (see Fig. 6*B*). The input from these CPn/PT cells to iMSNs is assumed to encode $I(S(t_{i-1}))$, and the activity of the iMSNs is assumed to be determined by imposing the input–output function f_2 on it, i.e., $f_2(I(S(t_{i-1})))$, which is equal to $V(S(t_{i-1}))$ under the condition in which f_2 operates in the suprathreshold linear regimen and is not affected by D_2 antagonist. We assume that individual subsets of CCS cells or CPn/PT cells project to different subsets of dMSNs or iMSNs, respectively, and define x_{dMSN} and x_{iMSN} as variables representing the population activity of all the subsets of dMSNs and iMSNs; because single subsets of CCS cells and CPn/PT cells are assumed to be active at each time step as described above, this results in $x_{dMSN} = f_1(I(S(t_i)))$ and $x_{iMSN} = f_2(I(S(t_{i-1})))$. If reward is obtained at a given state, the activity of the PPN neurons, x_{PPN}, is assumed to represent the reward value Rew, Rew = 10 (large-reward case) or Rew = 5 (small-reward case); otherwise, x_{PPN} is assumed to be 0. Given the activity of dMSNs, iMSNs, and the PPN neurons, dopaminergic neuronal response (deviation from the baseline activity), x_{DA}, is assumed to be determined by

$$x_{DA} = x_{PPN} + \gamma x_{dMSN} - x_{iMSN},$$

where γ represents the strength/efficacy of the direct pathway (relative to the indirect pathway) (see Fig. 6*C*). This indicates that the dopamine response represents

$$x_{DA} = \mathrm{Rew} + \gamma f_1(I(S(t_i))) - f_2(I(S(t_{i-1}))).$$

Under the condition in which the functions f_1 and f_2 operate in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists, this is equivalent to the following:

$$x_{DA} = \mathrm{Rew} + \gamma V(S(t_i)) - V(S(t_{i-1})),$$

which is equal to the TD reward prediction error for state value (Sutton and Barto, 1998), and γ corresponds to the time discount factor. Such dopaminergic response presumably induces plasticity in the corticostriatal (CCS–dMSNs and CPn/PT–iMSNs) connections, and it is assumed to be represented by a modification of $I(S(t_{i-1}))$:

$$I(S(t_{i-1})) \rightarrow I(S(t_{i-1})) + \alpha x_{DA},$$

where α is a constant representing the learning rate, and it was set to 0.75, the same as in the original simple model described above. Notably, this update (synaptic modification) is assumed to occur at every time step, not limited to the time point at which reward is obtained (as in the original model). The saccadic reaction time was assumed to be in the same quasi-inverse relationship with the activity of dMSNs at target presentation (i.e., at S_2 or S_3) as assumed for the simple model described above. The effects of D_1 and D_2 antagonists were also assumed to be the same as assumed in the simple model (see Fig. 3). The number of trials in individual blocks was set similarly to the case of the simple model described above. Both left-target and right-target trials were simulated, and Figure 6 shows the data for left-target trials [as for Fig. 6*D*,*E*, the first trials (shown in the left of the panels) were left-target trials and the second trials were left- or right-target trials (mixed)]; the model is left–right symmetric, and "left" and "right" are just labeling. Values of I were initialized to 0, and 521 task blocks were simulated in the absence of antagonists or with either D_1 and D_2 antagonist (the initial 10 blocks were simulated without antagonist in any cases) for each value of γ: $\gamma = 0.75$ and 0.9 (see Fig. 6) and 0.3 and 0.5 (data not shown). Average for the last 500 blocks are shown in Figure 6. As in the case of the original model, for the simulations, we wrote codes (driven equations) for 1/10 scaled-down values of the variables, and then the results were 10 times scaled up when they were illustrated.

As mentioned above, we ran simulations with different values of γ (strength/efficacy of the direct pathway relative to the indirect pathway). We described its purpose and results in detail below (see Results). Here we make another remark related to this point. As mentioned previously, there are direct projections from the striatal striosomes to the SNc dopamine neurons (Gerfen et al., 1985; Paladini et al., 1999; Fujiyama et al., 2011; Watabe-Uchida et al., 2012), but they are not incorporated into our model. At present, it is still somewhat ambiguous whether those projections are physiologically strong (Chuhma et al., 2011), and also the ratio with which the striosomal neurons express the D_1Rs or D_2Rs and as well as the ratio with which those cells receive inputs from CCS or CPn/PT cells are not clear (for the latest views on this issue, see Reiner et al., 2010; Crittenden and Graybiel, 2011). Our model would need to be revised when these things are clarified in the future. Notably, it has been shown previously (Fujiyama et al., 2011) that single striosomal neurons can project to both the SNr and SNc. If these neurons target GABAergic neurons (rather than dendrites of dopamine neurons) in the SNr and dopamine neurons in the SNc and also if they receive inputs predominantly from the CCS cells, existence of such co-projections could reduce the effective weight of the CCS–dMSNs–SNr–SNc pathway (i.e., CCS-direct pathway influence on the SNc) assumed in our model. In contrast, if the predominant upstream of the striosomal cells is the CPn/PT cells, existence of the co-projections to the SNr and the SNc would affect the presumed CPn/PT-indirect pathway influence on the SNc in our model. It has been shown previously (Watabe-Uchida et al., 2012) that not only the striosomal inputs but also dopamine neurons receive direct inputs from a number of structures, including the GP (rodent homolog of GPe). Any connections that were not incorporated into our model can potentially affect the effective value of the parameter γ (at the minimum, it is of course likely that those connections have certain functions other than modulating this parameter). As described in Results and shown in Figure 6, the elaborated model can well explain the observed distinct effects of the antagonists on reaction times if $\gamma = 0.75$ (actually, better than the original simple model; see Results) and could still explain them to a certain degree if $\gamma = 0.9$. We further confirmed that those effects could also be mostly reproduced if γ is set to 0.5 or 0.3 (data not shown). Therefore, at least in this sense, our model appears to have certain robustness against the unincorporated connections including the striosomal direct projections to the SNc.

Notably, the discrete time description of our model is certainly a very rough approximation, and construction of more detailed model with

continuous time is desired in the future. Nevertheless, some sort of temporal discreteness might in fact exist in the actual operation of the dopamine system, i.e., particular time points can be "marked" in a sense that dopamine release is specifically enhanced through, for example, slowly oscillating neural activity, in particular, the recently reported 4 Hz synchronized oscillation in the prefrontal cortex, VTA, and hippocampus (Fujisawa and Buzsáki, 2011), possibly coupled with cholinergic induction of dopamine release (Threlfell et al., 2012). A single time step (duration between states) in our model is assumed to be approximately up to a few (to several) hundred milliseconds, which seems broadly consistent with this possibility. This timescale could also be in line with a different line of experimental observation. Specifically, it has been shown that dopamine neurons sometimes respond to events that should have zero reward prediction error with a biphasic "excitation-then-inhibition" pattern (Kakade and Dayan, 2002) whose duration looks to be approximately a few hundred milliseconds. The biphasic response could reflect such a circuit architecture as the one assumed in our model, i.e., fast excitation via the CCS-direct pathway and delayed inhibition via the CPn/PT-indirect pathway, and the time interval from excitation to inhibition could represent a canonical time step of the system.

Model of a task involving action selection. The saccade task we modeled in the above consists of forced-choice trials only. To consider how our proposed circuitry operates in a situation in which subjects can voluntarily select an action based on their learned evaluations, we constructed an extended model of the corticobasal ganglia circuit incorporating a mechanism for action selection and simulated (a half of) the task involving action selection used in recent studies (Roesch et al., 2007; Takahashi et al., 2011); the latter study (Takahashi et al., 2011) has developed a computational model, and we extended our model in reference to their experimental results as well as the model. In the task, at the beginning of each trial, the subject (rat) was presented with one of three odor cues, two of which indicated that reward will be given if the animal entered either the left or the right well, respectively (i.e., forced choice), whereas the remaining cue indicated that the animal will be rewarded in both of the directions (i.e., free choice). In any case, the amount of reward, either small (one bolus of sucrose solution) or large (two boluses given sequentially), was determined according to the predetermined direction-amount contingency that was fixed during a block and then reversed. The three types of cues were presented according to a pseudorandom sequence so that the free-choice cue was presented in 7 of 20 trials and the two forced-choice cues were presented in equal numbers, and the same cue was not presented on more than three consecutive trials.

Takahashi et al. (2011) examined the effects of a lesion of the orbitofrontal cortex (OFC) on the dopamine neuronal activity and subjects' behavior. Comparing the results with a computational model developed within the work and also combining electrical stimulation, they have revealed that, in intact subjects, the OFC contains information about the "model" of the task structure, i.e., a diagram for transitions between different "states," and influences the VTA dopamine neurons so that they can compute reward prediction error according to the model. Based on these results and according to a demonstration of the existence of CCS cells and CPn/PT cells in the OFC (Hirai et al., 2012), we assumed that the CCS cells and CPn/PT cells in the OFC upregulate and downregulate VTA dopamine neurons, respectively, through the pathways shown in Figure 7*B*. At each time step (t_i), the subject is assumed to be in one of the states shown in Figure 9*A*, which is similar to (although not exactly the same as) the states proposed by Takahashi et al. (2011). There are either multiple options (at state S_2) or a single option (at the other states) that can be taken.

In the case with a single option (i.e., states other than S_2), a subset of CCS cells is assumed to represent the combination of that state and the available choice option $[A(t_i)]$, and the input from these CCS cells to dMSNs, denoted by $I(A(t_i))$, is assumed to represent the predicted value of option $A(t_i)$ [denoted by $Q(A(t_i))$], shifted by the amount of the threshold of the input–output function, i.e., $Q(A(t_i)) = I(A(t_i)) - \theta$. The activity of the dMSNs is assumed to be determined by imposing the input–output function f_1 on $I(A(t_i))$, i.e., $f_1(I(A(t_i)))$, which is equal to $Q(A(t_i))$ under the condition in which f_1 operates in the suprathreshold linear regimen and is not affected by D_1 antagonist. Meanwhile, a subset of CPn/PT cells is assumed to represent the combination of the state of the subject at the previous time step (t_{i-1}) and a choice option that has been taken at that state $[A(t_{i-1})]$. The input from these CPn/PT cells to iMSNs is assumed to encode $I(A(t_{i-1}))$, and the activity of the iMSNs is assumed to be determined by imposing the input–output function f_2 on it, i.e., $f_2(I(A(t_{i-1})))$, which is equal to $Q(A(t_{i-1}))$ under the condition in which f_2 operates in the suprathreshold linear regimen and is not affected by D_2 antagonist. We assume that individual subsets of CCS cells or CPn/PT cells project to different subsets of dMSNs or iMSNs, respectively, and define x_{dMSN} and x_{iMSN} as variables representing the population activity of all the subsets of dMSNs and iMSNs; because single subsets of CCS cells and CPn/PT cells are assumed to be active at each time step as described above, this results in $x_{\mathrm{dMSN}} = f_1(I(A(t_i)))$ and $x_{\mathrm{iMSN}} = f_2(I(A(t_{i-1})))$. If reward is obtained at a given state, the activity of the PPN neurons, x_{PPN}, is assumed to represent the reward value Rew (set to 10); otherwise, x_{PPN} is assumed to be 0. Specifically, $x_{\mathrm{PPN}} = \mathrm{Rew} = 10$ at S_8 or S_9 in the case of small reward, and $x_{\mathrm{PPN}} = \mathrm{Rew} = 10$ at (S_8 and S_{10}) or (S_9 and S_{11}) in the case of big reward.

In the state in which multiple options are available (i.e., in state S_2), we assumed that the "max" operation is implemented through the corticostriatal (CCS–dMSN) feedforward inhibition (Parthasarathy and Graybiel, 1997; Gittis et al., 2010) and possibly also through lateral or feedback inhibition, i.e., $x_{\mathrm{dMSN}} = f_1(\max\{I(A(t_i))\})$ so that the activity of dMSNs reflects the predicted value of an option that currently has the maximum predicted value $[\max\{Q(A(t_i))\}]$ under the condition in which f_1 operates in the suprathreshold linear regimen and is not affected by D_1 antagonist (see Fig. 8*C*). This assumption is based on a previous finding (Roesch et al., 2007) that the VTA dopamine neurons appear to represent TD reward prediction error for *Q*-learning (for a detailed explanation, see below and Results). The activity of iMSNs was assumed to be the same as the case with a single option described above.

Given the activity of dMSNs, iMSNs, and the PPN neurons, dopaminergic neuronal response (deviation from the baseline activity), x_{DA}, is assumed to be determined by

$$x_{\mathrm{DA}} = x_{\mathrm{PPN}} + \gamma x_{\mathrm{dMSN}} - x_{\mathrm{iMSN}},$$

where γ represents the strength/efficacy of the direct pathway (relative to the indirect pathway), and it was set to 0.75 in the simulations. This indicates that the dopamine response represents

$$x_{\mathrm{DA}} = \mathrm{Rew} + \gamma f_1(I(A(t_i))) - f_2(I(A(t_{i-1})))$$

in the case in which there is a single-choice option, and

$$x_{\mathrm{DA}} = \mathrm{Rew} + \gamma f_1(\max\{I(A(t_i))\}) - f_2(I(A(t_{i-1})))$$

in the case in which there are multiple options. Under the condition in which the functions f_1 and f_2 operate in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists, these are, respectively, equivalent to

$$x_{\mathrm{DA}} = \mathrm{Rew} + \gamma Q(A(t_i)) - Q(A(t_{i-1}))$$

and

$$x_{\mathrm{DA}} = \mathrm{Rew} + \gamma \max\{Q(A(t_i))\} - Q(A(t_{i-1})),$$

which equal to the TD reward prediction error for action value defined in *Q*-learning (Sutton and Barto, 1998), except for decay of plastic changes in the corticostriatal connection strengths that we also assumed as we describe below.

In the case in which there are multiple (two) choice options, i.e., at state S_2 (the two options are A_2 and A_3), one of the options were assumed to be selected according to a soft-max operation (see Fig. 9*A*, right graph) through competitive neural dynamics among the CPn/PT cells receiving the effects of dMSNs (Fig. 8*C*). Specifically, we assumed that the probabilities that options A_2 and A_3 are selected, $\mathrm{Prob}(A_2)$ and $\mathrm{Prob}(A_3)$, are given by

$$\mathrm{Prob}(A_2) = (1 + \exp(-\beta(f_1(I(A_2)) - f_1(I(A_3))))),$$

and

$$\mathrm{Prob}(A_3) = 1/(1 + \exp(-\beta(f_1(I(A_3)) - f_1(I(A_2))))) = 1 - \mathrm{Prob}(A_2),$$

respectively, where β is a parameter determining the balance between exploration and exploitation, and it was set to 0.5 in the simulations (for a detailed explanation on the presumed circuit operation, see Results). These are equal to

$$\mathrm{Prob}(A_2) = 1/(1 + \exp(-\beta(Q(A_2) - Q(A_3)))), \text{ and}$$

$$\mathrm{Prob}(A_3) = 1/(1 + \exp(-\beta(Q(A_3) - Q(A_2)))) = 1 - \mathrm{Prob}(A_2),$$

under the condition in which f_1 operates in the suprathreshold linear regime and is not affected by D1 antagonist.

Dopamine response is assumed to induce plasticity in the connections between CCS cells and dMSNs and those between CPn/PT cells and iMSNs at every time step, resulting in the following update rule:

$$I(A(t_i-1)) \rightarrow I(A(t_i-1)) + \alpha x_{\mathrm{DA}}.$$

$A(t_{i-1})$ represents an option that has actually been taken at time t_{i-1}, which is not necessarily the same as the one that has been once selected at the stage of dMSNs (i.e., the one with the maximum predicted value) because of the assignments of the hard-max and soft-max operations to the stages of dMSNs and CPn/PT cells, respectively, in our presumed implementation of Q-learning (see Fig. 8*C*). We propose that the subset of dMSNs corresponding to $A(t_{i-1})$, synapses on which should be modified according to x_{DA}, could be marked as the target of plasticity by receiving inputs from CPn/PT cells representing $A(t_{i-1})$ via CPn/PT–dMSN connections, which are minor but known to still exist (Lei et al., 2004; Reiner et al., 2010). Indeed, these minor inputs may cause only dendritic spikes/plateaus, which have been suggested to promote plasticity (Golding et al., 2002; Surmeier et al., 2011) as we mentioned above, but not somatic spikes so that they do not affect the computation of reward prediction error in the dopamine neurons or the action selection process. Notably, in the cases of our proposed implementations of SARSA (state–action–reward–state–action) (see Fig. 8*A*,*B*), mismatch between the maximally activated dMSNs and the eventually selected option never occurs. α is the learning rate and is set to 0.6. The reaction time was assumed to be quasi-inversely related to the activity of dMSNs at cue offset (at S_4 or S_5), in a manner similar to the assumption made in the models for the saccade task above:

$$RT = C_1/(C_2 + x_{\mathrm{dMSN}}),$$

with different parameters: C_1 and C_2 were set to 2300 and 10, respectively. The effects of D_1 and D_2 antagonists were assumed to be the same as assumed in the models for the saccade task (see Fig. 3).

We explored parameters of the model with which the performance of the model becomes comparable with the reported experimental data (Roesch et al., 2007; Takahashi et al., 2011). In the course of this exploration, we realized that an additional assumption seems to be necessary to be added to the model to reproduce important features of the neural activity data. In the experiment, the dopamine neuronal response to reward clearly remained and was in fact stronger than the response to a cue, even in the last 10 trials of a block (Roesch et al., 2007, their Fig. 4*d*). This indicates that, for the dopamine neurons (not necessarily for the animals), reward remained to be rather "surprising," not fully predictable, through an entire task block. At the same time, however, the dopamine neuronal response to a cue clearly increased or decreased depending on whether the cue was associated with large or small reward, respectively, indicating that the dopamine neuronal responses certainly reflect the progression of learning. To reproduce both of these features simultaneously, some sort of time-dependent decay of learned values seems to be required to be incorporated into the model, while the learning rate and the time discount factor, which are represented by the (relative) strength of the direct pathway over the indirect pathway in our model (Morita et al., 2012), should be kept reasonably high. We achieved this by assuming time-dependent decay of plastic changes of the corticostriatal connection strengths. Specifically, we made an additional assumption that changes in the corticostriatal connection strengths, or more precisely, deviations of $I(A_1)$, etc., from their initial (baseline) value, I_0, are subject to time-dependent decay at every time step:

$$I(A_i) \rightarrow I_0 + (I(A_i) - I_0) \times 0.9^{(1/16)}$$

(we assumed 16 time steps within a single task trial, and thus the above corresponds to the fact that changes in $I(A_i)$ decay to 90% per trial). I_0 was set to 4.5, which was below the threshold of the input–output functions of the MSNs in the case without antagonists ($\theta = 5$ as described above). We explored parameters with which the model qualitatively reproduced main features of the experimental results (Roesch et al., 2007; Takahashi et al., 2011) and obtained the values described in the above.

In a separate set of simulations, in one of the two wells (left and right), virtual optogenetic stimulation was applied to either dMSNs or iMSNs coincidently with a reward bolus (at S_8 or S_9), in addition to giving an extra bolus of reward at the subsequent timing in both of the wells. Specifically, in addition to setting Rew to 10 at S_{10} and S_{11}, which represents the extra bolus, the value of x_{dMSN} or x_{iMSN} at S_8 or S_9 was increased by 10, representing the optogenetic stimulation (10 was added directly to x_{dMSN} rather than to the input term, on which f_1 or f_2 was imposed, because currents induced by optogenetic stimulation would bypass the AMPA and NMDA channels and so would not be directly affected by the antagonists in the same way as the synaptic currents as we assumed in Fig. 3). The contingency between the stimulation on/off and the location of the well was fixed for a block and alternated across blocks in a similar manner to the contingency between the presence/absence of the extra bolus of reward and the location of the well in the original experiments.

The number of trials in each block was set to 120 [in the experiments by Roesch et al. (2007), it is described that at least 60 trials were collected per block]. The three types of cues were presented pseudorandomly, in a manner similar to the one in the experiment (Roesch et al., 2007) described above: the free-choice cue was presented on 7 of 20 trials, and the two forced-choice cues were presented in equal numbers, on average in each session, and the same cue was not presented on more than three consecutive trials. Values of I were initialized to $I_0 = 4.5$, and 2021 task blocks were simulated under seven conditions: (1) big versus small reward, without antagonists; (2) big versus small reward, with D_1 antagonist; (3) big versus small reward, with D_2 antagonist; (4) dMSN stimulation, without antagonists; (5) iMSN stimulation, without antagonists; (6) dMSN stimulation, with D_1 and D_2 antagonists; and (7) iMSN stimulation, with D_1 and D_2 antagonists. The averages for the last 2000 blocks are shown in Figure 9. As in the cases of the abovementioned models, for the simulations, we wrote codes (driven equations) for $1/10$ scaled-down values of the variables, and then the results were 10 times scaled up when they were illustrated.

Results

Corticobasal ganglia circuit for reward-oriented saccade

We first simulated a visually guided saccade task (Fig. 1*A*) used in an experimental study (Nakamura and Hikosaka, 2006). On each trial of the task, a visual target appeared at the left or the right of the screen, and the subject (monkey) was required to make a saccadic eye movement toward the target. If the subject made a correct saccadic response, a liquid reward was given after 100 ms of fixation at the target. There were two kinds of reward amount, large and small, each of which was associated with either the left or right target; contingency between the target location and the reward amount was fixed for an individual task block consisting of 20–28 trials, and left-large and left-small blocks were alternated. The target location (left and right) was pseudorandomly determined in each trial.

A series of studies have revealed that the cortex, basal ganglia, and the dopamine neurons in the SNc play essential roles in the learning and execution of reward-associated saccade tasks

(Kawagoe et al., 2004; Takikawa et al., 2004; Hikosaka et al., 2006; Nakamura and Hikosaka, 2006). We constructed a computational model of the corticobasal ganglia–SNc circuit according to a conjecture derived from recent anatomical and physiological findings (Morita et al., 2012) (Fig. 1*B*) (for details, see Materials and Methods). We assumed that there exist the two types of corticostriatal neurons, CCS cells and CPn/PT cells, in the saccade-related cortical areas, such as the FEF, and when a saccadic target is presented, its location is represented by a subset of CCS cells (Fig. 1*B*, left). These CCS cells activate dMSNs, which then inhibit neurons in the SNr and thereby disinhibit the superior colliculus (SC) neurons to initiate a saccade (Kawagoe et al., 2004; Hikosaka et al., 2006). Meanwhile, the CCS cells activate a subset of CPn/PT cells through the unidirectional projections (Morishima and Kawaguchi, 2006). These CPn/PT cells presumably sustain their activity (Morita et al., 2012) via the strong recurrent excitation (Morishima et al., 2011), whereas the activity of the CCS cells presumably declines because recurrent excitation among them is weaker and depressive. When reward is obtained (Fig. 1*B*, right), dopamine neurons in the SNc are assumed to receive excitatory inputs from the PPN neurons, which presumably inform the value (amount) of the obtained reward (cf. Okada et al., 2009). The dopamine neurons also receive inhibitory inputs from the collaterals of SNr projection neurons (Tepper and Lee, 2007), which are disinhibited by the iMSNs downstream of the still active CPn/PT cells via the indirect pathway. Given that the activity of iMSNs represents the reward value (reward amount) predicted from the target location represented now by the CPn/PT cells ["chosen value" (cf. Lau and Glimcher, 2008); see below], the dopaminergic neuronal response represents a difference between the obtained and the predicted reward values, i.e., reward prediction error (Morita et al., 2012). The phasic dopamine response is then assumed to induce proportional plastic changes of the synaptic strengths between CCS cells and dMSNs (cf. Reynolds et al., 2001) so that the activity of dMSNs can represent updated reward value predictions (cf. Samejima et al., 2005; Lau and Glimcher, 2008). At the synapses between CPn/PT cells and iMSNs, sustained inputs from CPn/PT cells presumably cause adenosine accumulation (cf. Cunha, 2001; Schiffmann et al., 2007) and glutamate spillover (cf. Mitrano et al., 2010). Phasic increase in dopamine would then stimulate the signaling cascade downstream of the adenosine A_{2A} receptors (cf. Azdad et al., 2009), and it is assumed to lead to LTP, whereas phasic decrease in dopamine is assumed to cause LTD through the signaling cascade downstream of mGluR5 (for details, see Materials and Methods). In consequence, the activity of iMSNs can also represent updated reward value predictions, effectively in the same manner as the activity of dMSNs.

For simplicity, and in the same manner as the previous modeling study (Hong and Hikosaka, 2011), we simulated only the trials with a left target with our first model; later we simulated all the trials with an elaborated model but show the results for left-target trials unless otherwise described. We calculated the activity of dMSNs, iMSNs, and SNc dopamine neurons for 500 task blocks. Notably, we have not explicitly modeled structures between the striatum and the dopamine neurons, including the SNr, and our model can be also compatible with recent findings (Matsumoto and Hikosaka, 2007; Hong and Hikosaka, 2008; Hong et al., 2011; Shabel et al., 2012) that the dopamine neurons are regulated by the LHb neurons, which are then driven by GPb (see Figs. 7*A*,*B*).

Behavioral modulation by reward amount and underlying dopaminergic neuronal response

Figure 2*A* shows the stimulus-induced response of striatal dMSNs downstream of the cortical CCS cells representing the left-target location, averaged over trials in which the target was presented in the left and aligned at the switches of blocks from left-large reward to left-small reward and from left-small to left-large. As shown in the figure, the response of dMSNs is greater in the blocks in which the left target is associated with large reward than in the blocks with small reward, accompanying gradual changes after the switch of the two conditions. Because we assumed that the activity of the CCS cells (upstream of dMSNs) is transmitted to the CPn/PT cells (upstream of iMSNs) and kept in them until reward is obtained (Fig. 1*B*, right), the activity of the iMSNs at the timing of reward reception shows the same pattern as the stimulus-induced response of the dMSNs (Fig. 2*B*). The SNc dopamine neurons are assumed to be inhibited by the iMSNs through the indirect pathway and the SNr, while receiving excitatory inputs informing the obtained reward value from the PPN. Consequently, their response pattern (Fig. 2*C*) is upside-down of the response pattern of iMSNs (Fig. 2*B*) shifted by the block-dependent reward amounts.

In fact, the response pattern of the dopamine neurons (Fig. 2*C*) also matches the mathematical differential (in an approximate sense) of the response pattern of the MSNs (Fig. 2*A*,*B*), and, in turn, the response pattern of MSNs matches the mathematical integral of the dopaminergic neuronal response pattern. This is because the phasic dopamine response is assumed to induce proportional changes of the corticostriatal connection strength. The fact that the response pattern of the dopamine neurons can be explained by both the subtraction of the pattern of the iMSNs from the block-dependent constant (representing the PPN inputs) and the differential of the pattern of the iMSNs (or dMSNs) is indicative of the fact that the entire system can operate autonomously without assuming the response pattern of inputs from unspecified upstream to the dopamine neurons. Functionally, these relationships can be interpreted as follows: (1) the activity of the MSNs represents predicted reward value; (2) the dopamine neurons subtract the iMSN activity from the PPN activity representing the obtained reward value, so as to compute reward prediction error; and (3) this error signal is used to update the predicted reward value stored in the corticostriatal connections. The trial-by-trial gradual changes in the dopaminergic neuronal response predicted from our model (Fig. 2*C*) well match those experimentally observed during a similar task (Takikawa et al., 2004) (Fig. 2*D*).

In reference to empirical findings (Kawagoe et al., 2004) and the results of the previous modeling study (Hong and Hikosaka, 2011), we assumed that the level of the stimulus-induced response of dMSNs is quasi-inversely related to the saccadic reaction time (see Materials and Methods). Such a relationship is expected to hold, given that the dMSNs inhibit the SNr neurons and thereby disinhibit the SC neurons to initiate a saccade (Fig. 1*B*, left). With this assumption, our model predicts shorter reaction time in blocks with large reward than in small-reward blocks (Fig. 2*E*), reproducing the experimental results (Nakamura and Hikosaka, 2006).

Motivational effects of dopamine receptor antagonists

Having seen how the model operates under the normal condition, let us now consider the effects of dopamine receptor antagonists. Dopamine modulates the responsiveness of MSNs to cortical inputs through its tonic concentration and controls plas-

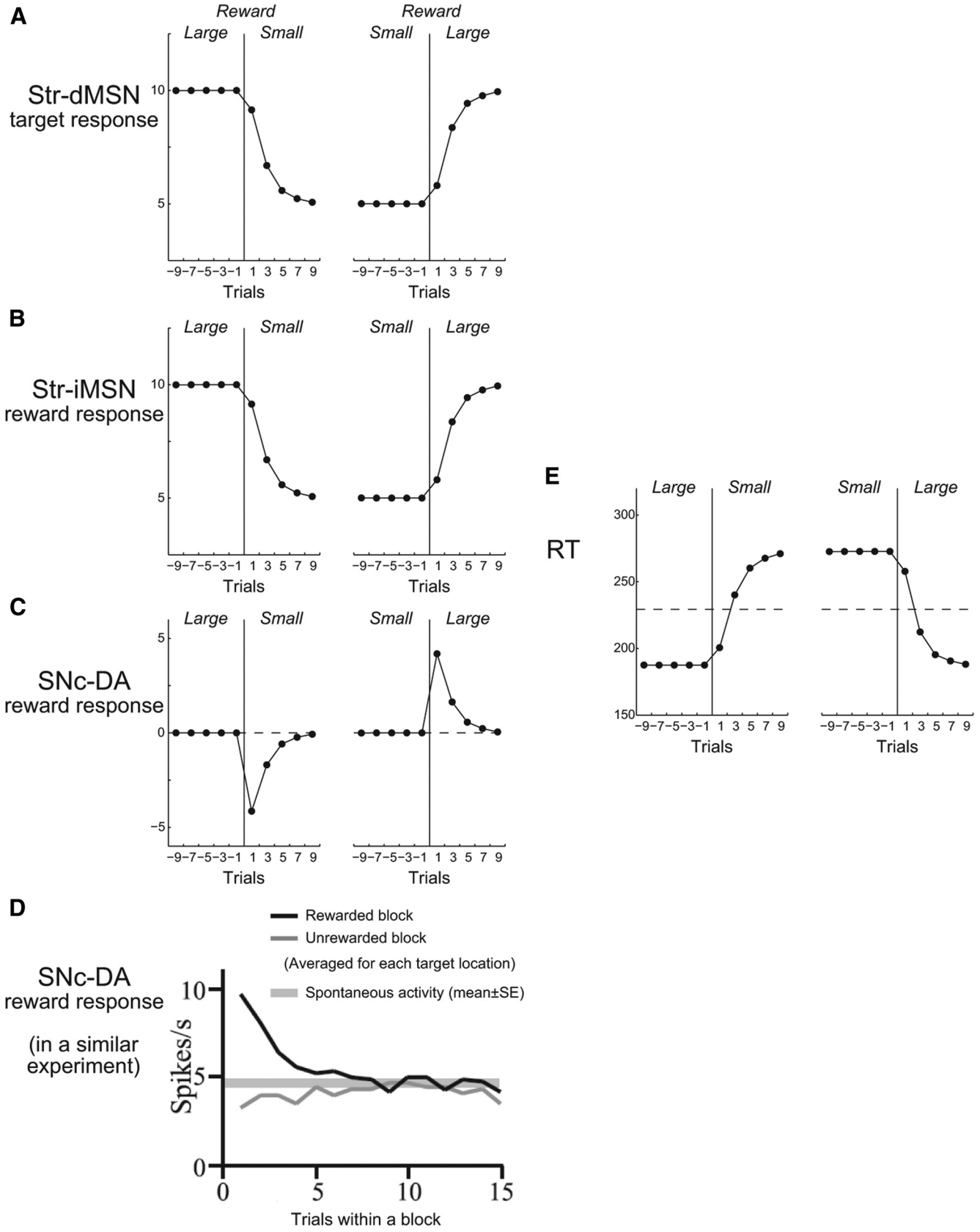

Figure 2. Activity of neurons and reaction times predicted from the model. ***A***, Stimulus-induced response of striatal (Str) dMSNs downstream of the cortical CCS cells representing the left-target location, aligned at the switch of blocks from left-large reward to left-small reward and from left-small to left-large. ***B***, Response of iMSNs at the timing of reward reception. It exhibits the same pattern as the stimulus-induced response of dMSNs because the activity of the CCS cells (upstream of the dMSNs) is presumably transmitted to the CPn/PT cells (upstream of the iMSNs) and sustained there (Fig. 1*B*, right). ***C***, Response of SNc dopamine neurons at the timing of reward reception, which is presumably determined by the excitatory inputs from the PPN neurons representing the obtained reward value (large or small) and the net inhibitory inputs from the iMSNs (through the indirect pathway and the SNr) representing the predicted reward value; (*Figure legend continues.*)

ticity induction in the corticostriatal synapses through its phasic changes (Gerfen and Surmeier, 2011). Both of these effects can potentially be blocked by dopamine receptor antagonists. However, here we concentrate on the blockade of the tonic dopaminergic modulation of the responsiveness of the MSNs, because phasic change in dopamine is expected to have larger amplitude than the change in the tonic level so that it could escape from the blocking effect of antagonists at least to a certain extent. This assumption could potentially be supported by the finding (Pennartz et al., 1993) that application of certain concentrations of dopamine receptor antagonists in the ventral striatum slices did not significantly change the amount of LTP induced by tetanic stimulation, which could cause phasic release of dopamine from the dopaminergic axonal fibers. Notably, our assumption is opposite from that of the previous modeling study (Hong and Hikosaka, 2011), which did not consider the effect of antagonists on the responsiveness of the MSNs but considered the effect on the synaptic plasticity.

Regarding the tonic dopaminergic modulation of the responsiveness of the MSNs, it is known that dopamine upregulates and downregulates the response of dMSNs and iMSNs via the activation of D_1Rs and D_2Rs, respectively (Gerfen and Surmeier, 2011). More precisely, it has been suggested that D_1R activation primarily enhances the NMDA current (Levine et al., 1996; Flores-Hernández et al., 2002; Moyer et al., 2007). Because the NMDA receptors have voltage dependence (Schiller and Schiller, 2001), the NMDA current, and its enhancement by D_1R activation, would become prominent when the dMSNs receive strong glutamatergic cortical (CCS) inputs. We thus assumed that the response of dMSNs to strong inputs, which would normally be enhanced by D_1R activation, is reduced by D_1 antagonist (Fig. 3*A*). Conversely, existing evidence suggests that the activation of D_2Rs normally suppresses the AMPA current and/or the NMDA receptor-mediated excitation (Moyer et al., 2007; Azdad et al., 2009; Higley and Sabatini, 2010), but the effect on the latter is blocked when A_{2A} adenosine receptors, which also exist in the iMSNs, are simultaneously activated (Azdad et al., 2009; Higley and Sabatini, 2010). Because adenosine is suggested to be synaptically generated from ATP released with neurotransmitter (e.g., glutamate) from presynaptic cortical cells (Cunha, 2001; Schiffmann et al., 2007), the suppressive effect of D_2R activation would be blocked by adenosine when the iMSNs receive strong cortical (CPn/PT) inputs and there exists a high amount of adenosine. Based on these considerations, we assumed that the response of iMSNs to weak inputs, which would normally be suppressed by D_2R activation, is enhanced by D_2 antagonist (Fig. 3*B*) (for details, see Materials and Methods). Figure 4 shows the effects of D_1 antagonist on the neuronal activity and the saccadic reaction time (Fig. 4*A–D*, black and red lines indicate the conditions without and with D_1 antagonist, respectively; these two are overlapped in Fig. 4*B*,*C*) compared with the experimental results (Nakamura and Hikosaka, 2006) (Fig. 4*E*). As shown in Figure 4*A*, the stimulus-induced response of dMSNs in large-reward blocks is reduced by D_1 antagonist. This is a direct outcome of the assumed reduction of the response of dMSNs to strong inputs (Fig. 3*A*). In contrast, D_1 antagonist does not affect the activity of iMSNs (Fig. 4*B*), which presumably do not express D_1Rs (Gerfen and Surmeier, 2011). The activity of dopamine neurons at the timing of reward is also not affected by D_1 antagonist (Fig. 4*C*), because in the model we now consider (Fig. 1*B*, right), it is determined solely by the inputs from the PPN and iMSNs, both of which are presumably not affected by D_1 antagonist (later we will consider effects of possible inputs from dMSNs using an elaborated model). Because we assumed that the reaction time is quasi-inversely related to the stimulus-induced response of the dMSNs, D_1 antagonist slows a reaction to stimulus in large-reward blocks (Fig. 4*D*), consistent with the experimental results (Nakamura and Hikosaka, 2006) (Fig. 4*E*). Figure 5 shows the effects of D_2 antagonist (Fig. 5*A–D*, black and blue lines indicate the conditions without and with D_2 antagonist, respectively) compared with the experimental results (Nakamura and Hikosaka, 2006) (Fig. 5*E*). This time, the stimulus-induced response of dMSNs in small-reward blocks is decreased in the presence of D_2 antagonist (Fig. 5*A*). However, we assumed that D_2 antagonist does not affect the responsiveness of dMSNs, because D_2R is presumably not expressed in the dMSNs (Gerfen and Surmeier, 2011). Therefore, such a decrease of the dMSNs response should not be a direct effect of D_2 antagonist. Conversely, D_2 antagonist was assumed to enhance the response of iMSNs to weak inputs (Fig. 3*B*), and, reflecting this, the activity of iMSNs is increased in the presence of D_2 antagonist in small-reward blocks (Fig. 5*B*, indi-

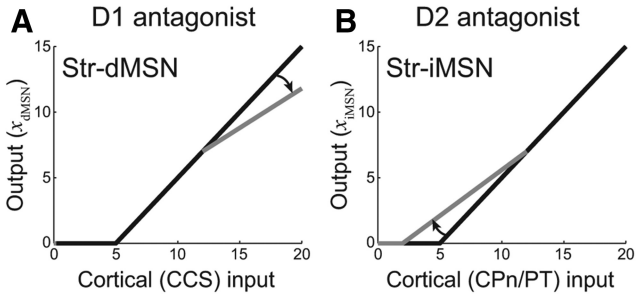

Figure 3. Presumed effects of dopamine receptor antagonists on the responsiveness of MSNs. ***A***, The black line represents the relationship between the strength of inputs from CCS cells to dMSNs (horizontal axis) and the response of dMSNs (vertical axis) in the presence of tonic dopamine. Response of dMSNs to strong cortical (CCS) inputs, which would be mediated significantly by the NMDA receptor current that is suggested to be enhanced by D_1R activation, was assumed to be attenuated by the application of D_1 antagonist (arrow and gray line). ***B***, The black line represents the relationship between the strength of the inputs from CPn/PT cells to iMSNs (horizontal axis) and the response of iMSNs (vertical axis) in the presence of tonic dopamine. Response of iMSNs to weak cortical (CPn/PT) inputs, which would normally be downregulated by D_2R activation through the suppression of the AMPA and/or NMDA receptor-mediated excitation, was assumed to be enhanced by the application of D_2 antagonist (arrow and gray line). Response of iMSNs to strong cortical (CPn/PT) inputs, which would normally be predominated by NMDA components, was assumed not to be enhanced by D_2 antagonist, considering that strong cortical inputs would cause generation of a high amount of adenosine, which presumably blocks the suppressive effect of D_2R activation on the NMDA receptor-mediated excitation. Notably, the actual relationship between the input strength and the output firing rate in the case without antagonists may entail saturating nonlinearity, but we assumed the threshold–linear function for simplicity. Str, Striatum.

←

(*Figure legend continued.*) it thus represents reward prediction error. Notably, the pattern of dopaminergic neuronal response also matches the mathematical differential (in an approximate sense) of the response pattern of the MSNs (***A***, ***B***), or, in turn, the pattern of the MSNs matches the mathematical integral of the dopaminergic neuronal pattern, because the corticostriatal (i.e., CCS– dMSN and CPn/PT–iMSN) connection strengths are assumed to be changed in proportion to the dopaminergic neuronal response. ***D***, Experimentally observed dopaminergic neuronal response at the timing of reward (Takikawa et al., 2004) during a slightly different task, in which one of four (rather than 2) possible target locations was associated with reward in each block (thus, there were twice more unrewarded → unrewarded switches than rewarded → unrewarded switches for each location, and this possibly explains the smaller amplitude of the change of the activity at the beginning of the unrewarded blocks than the change at the beginning of the rewarded blocks). This panel was taken from Takikawa et al. (2004) with modification. ***E***, Predicted saccadic reaction time, which was assumed to be quasi-inversely related to the stimulus-induced response of the dMSNs.

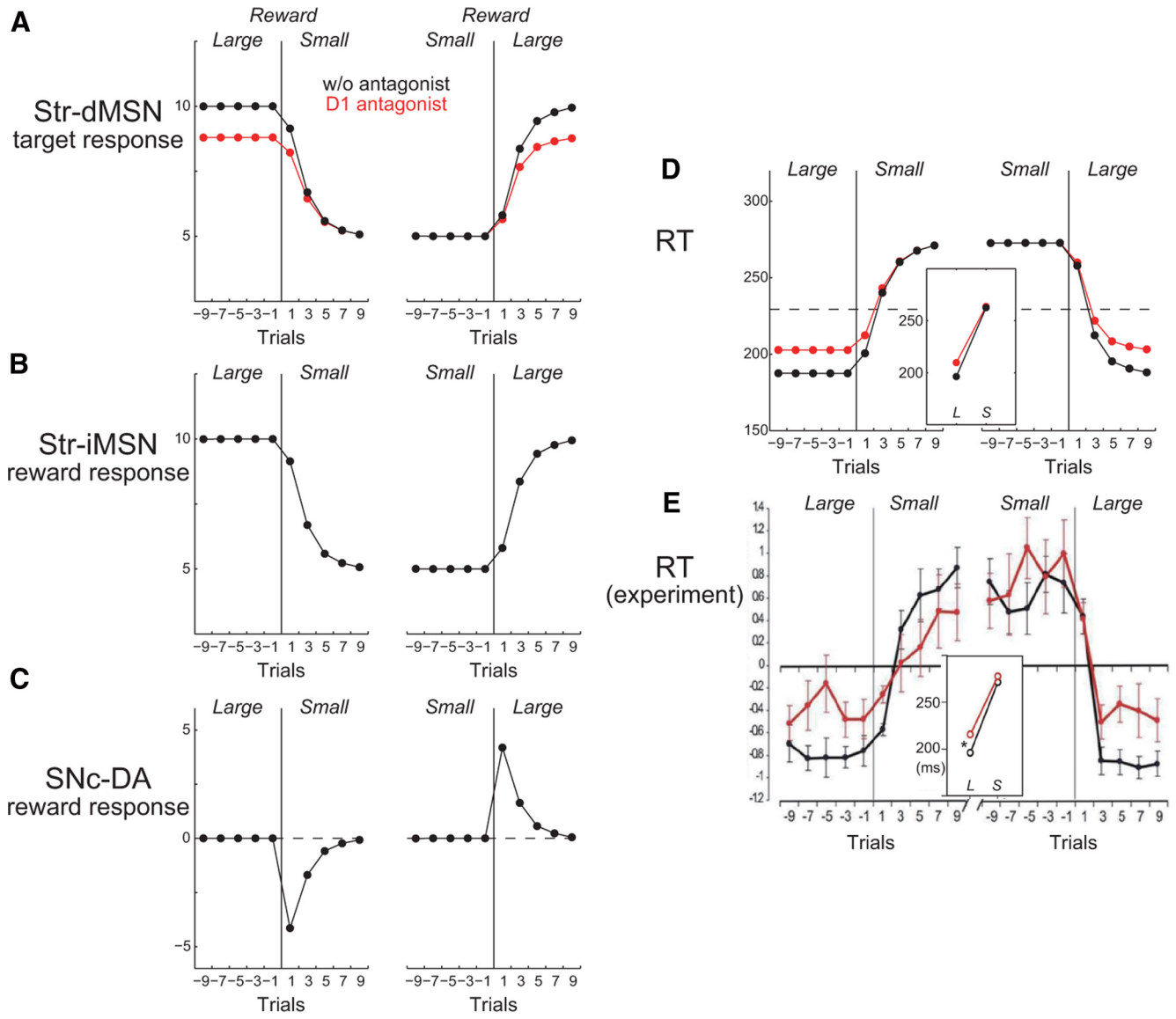

Figure 4. Predicted effects of D_1 antagonist on the neuronal activity and reaction time. ***A–D***, The black and red lines indicate the simulated neuronal response (***A–C***) and the saccadic reaction time (***D***) in the conditions without and with D_1 antagonist, respectively; these two are overlapped in ***B*** and ***C***. Notations are the same as those in Figure 2*A–C* and *E*. The stimulus-induced response of dMSNs in large-reward blocks is reduced by D_1 antagonist (***A***) because of the presumed suppressive effect of D_1 antagonist when dMSNs receive strong inputs (Fig. 3*A*), whereas the activity of iMSNs is not affected (***B***) because iMSNs do not express D_1Rs. The activity of dopamine neurons at the timing of reward reception is also not affected by D_1 antagonist (***C***), because it is presumably determined by the inputs from the iMSNs-indirect pathway and the PPN; the activity of dMSNs is assumed to already decline at this timing (Fig. 1*B*, right) and thus have little effect (however, other dMSNs may have activity representing expectation of rewards in the following trials, and that possibility is examined in the elaborated model; Fig. 6). Given that the reaction time was assumed to be quasi-inversely related to the stimulus-induced response of dMSNs, D_1 antagonist slows a reaction to stimulus in large-reward blocks (***D***). The inset shows the average reaction times for large- and small-reward blocks. ***E***, Experimental results for the effect of D_1 antagonist on the reaction time (Nakamura and Hikosaka, 2006). The black and red lines indicate the conditions without and with D_1 antagonist, respectively, in the same manner as in ***D***; this panel was taken from Hong and Hikosaka (2011). Str, Striatum.

cated by blue open triangles). However, such an increase almost disappears in the later part of the block, again indicating the existence of some indirect effects. Because the dopaminergic neuronal response at the timing of reward reception is assumed to be a subtraction of the activity of iMSNs from the excitatory PPN inputs, it entails the same pattern of the effect of D_2 antagonist but upside-down (Fig. 5*C*, blue open triangles).

As mentioned previously, we assumed that the strength of corticostriatal connections (both CCS–dMSNs and CPn/PT–iMSNs) is plastically changed according to phasic dopamine response, implementing the update of reward value prediction. Therefore, the decrease in the response of dopamine neurons in the presence of D_2 antagonist (Fig. 5*C*) causes a decrease in the strength of the corticostriatal connections. Iterative operations of this over successive trials then cause an accumulative decrease in the strength of the synapses between the CCS cells and dMSNs and thereby a decrease in the stimulus-induced response of dMSNs; this is what we observed in Figure 5*A* (indicated by blue filled triangles). The same accumulative decrease also occurs in the strength of the synapses between the CPn/PT cells and iMSNs, and it can explain why the increase in the iMSN activity seen in the early part of small-reward blocks (Fig. 5*B*, blue open triangles) disappears in the later part of the block. As such, the iMSNs receive two types of effects of D_2 antagonist: (1) direct facilitation, which operates immediately; and (2) indirect effect in the form of decrease in the input from the cortex (CPn/PT cells),

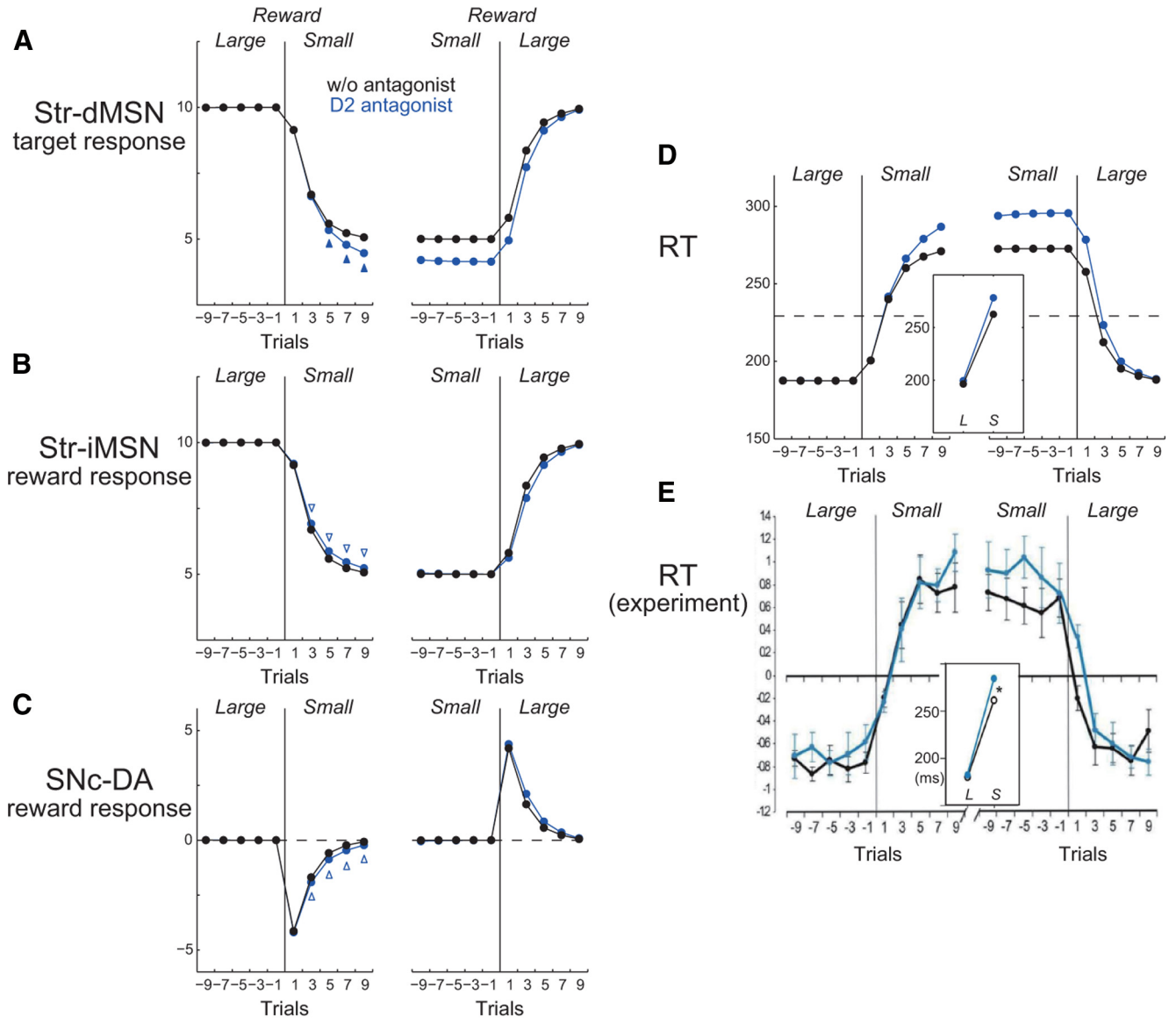

Figure 5. Predicted effects of D_2 antagonist. ***A–D***, The black and blue lines indicate the conditions without and with D_2 antagonist, respectively. Although D_2 antagonist does not have direct effect on dMSNs, it causes an increase in the activity of iMSNs in small-reward blocks (***B***, blue open triangles) because of the presumed enhancement of their response to weak cortical inputs (Fig. 3*B*). Because the dopaminergic neuronal response to reward is presumably a subtraction of the iMSN activity from the PPN inputs representing the obtained reward value, it entails the same pattern of the effect of D_2 antagonist but upside-down (***C***, blue open triangles). This decrease in the dopaminergic neuronal response causes a decrease in the corticostriatal (i.e., CCS– dMSN and CPn/PT–iMSN) connection strength through the presumed plasticity mechanism, resulting in an accumulative decrease in the input to, and the activity of, the dMSNs (***A***, blue filled triangles). The same accumulative decrease in the input also occurs in the iMSNs, gradually canceling out the direct facilitating effect of D_2 antagonist, explaining why the increase in iMSN activity seen in the early part of small-reward blocks (***B***, blue open triangles) disappears later. Because the reaction time was assumed to be quasi-inversely related to the stimulus-induced response of dMSNs, application of D_2 antagonist manifests as an increase in the reaction time in small-reward blocks (***D***). The inset shows the average reaction times for large- and small-reward blocks. ***E***, Experimental results for the effect of D_2 antagonist (Nakamura and Hikosaka, 2006); this panel was taken from Hong and Hikosaka (2011). Str, Striatum.

which operates with delay. Because we assumed that the reaction time is quasi-inversely related to the stimulus-induced response of dMSNs, application of D_2 antagonist manifests as an increase in the reaction time in small-reward blocks (Fig. 5*D*), well explaining the experimental results (Nakamura and Hikosaka, 2006) (Fig. 5*E*).

Consideration on reward expectation over multiple trials

In the saccade task (Nakamura and Hikosaka, 2006) that we simulated, reward was given just once in each trial, after the subject made a correct saccadic response, and no further reward was given until the next trial started. With this in mind, so far we assumed that the activity of dMSNs at the timing of reward, which represents the predicted value of the "end-of-trial" state in the model, was 0 as mentioned above. However, because the next trial started only a few seconds later and it has been shown (Enomoto et al., 2011) that the response of dopamine neurons can in fact reflect expectation of multiple rewards, our assumption may not be very appropriate. To more accurately simulate the real situation and examine whether our main results can still hold, we extended our model to incorporate reward expectation over multiple trials (Fig. 6*A–C*). Specifically, we assumed that the subject experienced transitions of a sequence of states (Fig. 6*A*), which include both within-trial states, at the target and reward timings, and multiple internal states corresponding to time epochs during in-

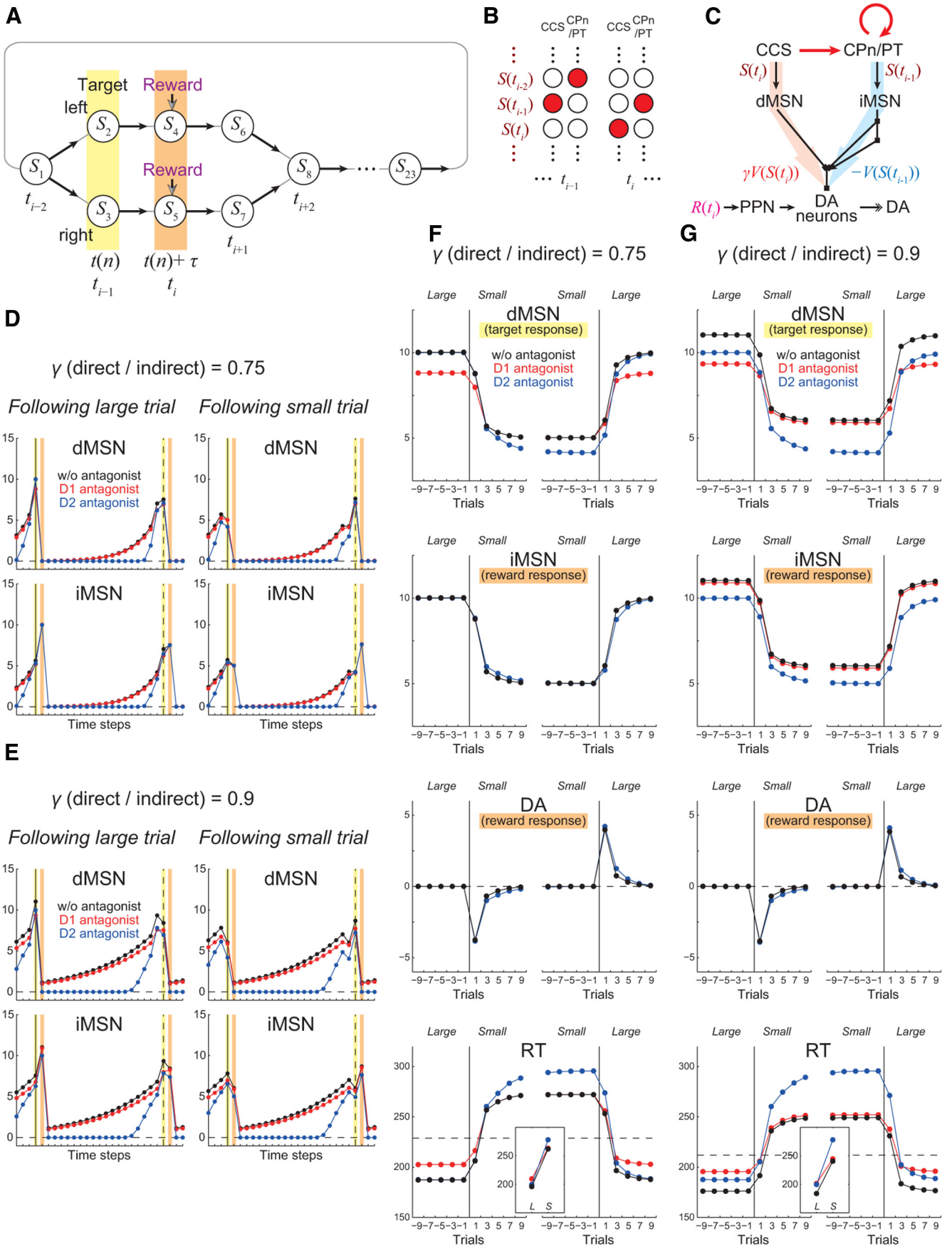

Figure 6. Consideration on reward expectation over multiple trials. ***A–C***, Elaborated model for the saccade task. ***A***, Presumed state transitions. ***B***, It is assumed that, while the subject experiences these state transitions, at each state, a subset of CCS cells actively represents that state (current state) and a subset of CPn/PT cells actively represents the previous (*Figure legend continues.*)

tertrial intervals. At each state, a subset of cortical CCS cells and CPn/PT cells are assumed to represent that state (current state) and the previous state, respectively (Fig. 6*B*), and dMSNs or iMSNs, receiving inputs from the CCS cells and CPn/PT cells, respectively, are assumed to represent the predicted values of those states (for details, see Materials and Methods). The dopamine neurons are then assumed to receive net excitatory and inhibitory effects from the dMSNs and iMSNs, respectively, and also receive excitatory inputs from the PPN neurons when reward is obtained (Fig. 6*C*) so as to compute TD reward prediction error that is presumably used to plastically modify the CCS–dMSN and CPn/PT–iMSN connection strengths, in the same manner as we assumed in the original simple model.

We ran simulations using this new model, varying the relative strength/efficacy of the direct pathway over the indirect pathway. In our model, this parameter, γ, represents a degree to which evaluation of future rewards is discounted depending on time distance, namely, time discount factor (Fig. 6*C*) (Morita et al., 2012). The black lines in Figure 6, *D* and *E*, show the activity of MSNs during two successive trials and intertrial intervals between them in the cases with $\gamma = 0.75$ and $\gamma = 0.9$, respectively (these should be regarded as the population activity; for details, see Materials and Methods): given that γ represents time discount factor for a single time step (approximately a few hundred milliseconds), these settings at 0.75 to 0.9 are overlapped with the range of experimentally measured values of discount factor for a longer duration for early and advanced learning stages (Enomoto et al., 2011). As shown in the figures, dMSNs and iMSNs show buildup activity during intertrial intervals, which presumably represent expectation of reward in the following trial. Such buildup activity is shaped through dopamine-dependent plasticity at the beginning of the task, with the "front" of the buildup activity shifted backward (in time) over trials and blocks (data not shown). Notably, there is a difference depending on the parameter γ. If γ is moderate ($\gamma = 0.75$), there appears to be no MSN activity at the beginning of an intertrial interval (i.e., end of a trial) (Fig. 6*D*, black lines). This indicates that the subject would have no expectation of future reward at the end of a trial, or in other words, the predicted value of the end-of-trial state is 0. In contrast, when the relative strength/efficacy of the direct pathway (time discount factor) was set to a higher value ($\gamma = 0.9$; Fig. 6*E*, black lines), the dMSNs and iMSNs come to show activity at the end of a trial, indicating that subject now has reward expectation beyond a single trial.

We examined the effects of dopamine receptor antagonists in this new model. In the case with the moderate γ ($\gamma = 0.75$), both D_1 and D_2 antagonists caused essentially the same effects as those observed in our original model, on the neural activity as well as on the reaction time (Fig. 6*F*). Looking more precisely, the sharp change in the reaction time on a transition from small- to large-reward blocks and the effect of D_1 antagonist on it (Fig. 4*E*, rightmost part of the panel) are now reproduced much better (Fig. 6*F*, bottom) than in the original simple model (Fig. 4*D*). As the parameter γ increases to be 0.9, however, there emerges another deviation from the original model (Fig. 6*G*). Specifically, whereas the reward-amount specificity of the effect of D_1 antagonist (i.e., prominent effect in large-reward trials) is essentially unchanged, D_2 antagonist comes to affect not only the neural activity and reaction time in small-reward trials but also those in large-reward trials, although less prominently.

The reason for this can be understood by looking at the effect of D_2 antagonist on the buildup activity of dMSNs and iMSNs during intertrial intervals (Fig. 6*E*, compare the black and blue lines). As shown in the figures, D_2 antagonist significantly attenuates the buildup activity, whereas D_1 antagonist has only a miner effect. Application of D_2 antagonist, which was done after 10 blocks have been completed without antagonists in the simulations using the elaborated model [antagonist application was preceded by no-antagonist blocks also in the experiment (Nakamura and Hikosaka, 2006)], initially enhances the buildup activity of iMSNs (data not shown), because inputs to the iMSNs during intertrial intervals are weak enough to be affected by D_2 antagonist (Fig. 3*B*). This causes negative response of dopamine neurons, resulting in a decrease of corticostriatal connection strengths and eventually the degradation of the buildup activity (of both dMSNs and iMSNs) that represents the expectation of reward in the following trial. Because such over-trial reward expectation is added on the value of the state at target presentation represented by dMSNs, which presumably determines the reaction time, regardless of the amount of reward associated with the target, its impairment by D_2 antagonist affects reaction time in both large-reward trials and small-reward trials. Despite this, however, the effect of D_2 antagonist on the reaction time is still larger for small-reward trials than for large-reward trials (Fig. 6*G*, bottom, inset), presumably because the mechanism explained previously (Fig. 5) should also operate. This indicates that our model with $\gamma = 0.9$ could still explain the observed reward-amount specificity of the effects of the antagonists (Nakamura and Hikosaka, 2006) at least to a certain degree.

Reward prediction error computation in parallel with action selection and/or execution

As we have so far shown, our circuit model can simultaneously explain the dopaminergic control of reaction time for a reward-

←

(*Figure legend continued.*) state (red and white circles in the figure indicate active and inactive subsets, respectively). ***C***, Receiving inputs from the active CCS cells and CPn/PT cells, dMSNs and iMSNs presumably represent the predicted value of the current and previous states, respectively, under the condition in which the input–output functions f_1 and f_2 of MSNs operate in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists. The dopamine neurons are assumed to receive net positive and negative effects from the dMSNs and iMSNs, respectively, and also reward-representing inputs from the PPN when reward is obtained, and thereby compute TD reward prediction error, which is then used to plastically modify the corticostriatal (i.e., CCS–dMSNs and CPn/PT–iMSNs) connection strengths. In our model, the relative strength/efficacy of the direct pathway over the indirect pathway (γ) represents the time discount factor appeared in the definition of the TD reward prediction error in reinforcement learning algorithms. ***D***, ***E***, Population activity of dMSNs and iMSNs during two successive trials and intertrial intervals between them, for two different values of γ (***D***, $\gamma = 0.75$; ***E***, $\gamma = 0.9$). The black, red, and blue lines indicate the cases without antagonists, with D_1 antagonist, and with D_2 antagonist, respectively. The left and right columns show the average for a duration after a large-reward trial (intertrial interval + next trial) and those after a small-reward trial, respectively. The vertical solid lines (with light-yellow backgrounds) indicate the timing of target presentation in the large-reward (left column) or small-reward (right column) trials, and the vertical dashed lines (with light-yellow backgrounds) indicate target presentation in the next trials (for which large- and small-reward trials were mixed). Light-orange backgrounds indicate the timings for reward. The initial 10 trials in each block are not included in the average so as to see that the situation after learning of location–reward contingency essentially reaches a steady state. Both left-target and right-target trials were simulated, and the data with the first trials (in the left of the panels) being left-target trials were shown here. ***F***, ***G***, Response of dMSNs at the timing of target presentation (top row), responses of iMSNs and the dopamine neurons at the timing of reward (middle 2 rows), and the reaction time (bottom row) in simulations with the two different values of γ (***F***, $\gamma = 0.75$; ***G***, $\gamma = 0.9$). The black lines indicate the cases without dopamine receptor antagonists, whereas the red and blue lines indicate the cases with D_1 and D_2 antagonist, respectively. Notations are the same as those in Figure 4 and 5. Left-target trials were shown in these panels (the model is left-right symmetric, and "left" and "right" are just labeling).

associated target, including the results of the pharmacological manipulations, and the dopaminergic representation of reward prediction error. However, because the task that we simulated (Nakamura and Hikosaka, 2006) included forced-choice trials only, we could not show whether our model can also operate well in a situation in which multiple-choice options are available and the reward prediction error signal is actively used for biasing advantageous actions. To address this issue, we simulated a different task used in different studies (Roesch et al., 2007; Takahashi et al., 2011) (precisely speaking, half of the task used in these studies). This task regards rats' movements rather than monkeys' saccade, but both tasks share several common features: (1) there are two action directions (left and right), which are associated with either big or small reward, and (2) the contingency between the action directions and the amount of reward is alternated in a blockwise manner. However, there is an important difference: specifically, the task used in the latter works (Roesch et al., 2007; Takahashi et al., 2011) included free-choice trials interleaved in forced-choice trials. By incorporating this feature, the authors (Roesch et al., 2007) succeeded in examining the activity of dopamine neurons in the course of active instrumental learning and found that the dopamine neurons appear to represent a specific type of reward prediction error that is defined in a particular reinforcement learning algorithm, namely, *Q*-learning (Watkins, 1989), rather than those defined in other algorithms, such as SARSA (Rummery and Niranjan, 1994). Moreover, in a subsequent study using the same task (Takahashi et al., 2011), the authors examined the effects of a lesion of the OFC, in combination with computational modeling and electrical stimulation, and revealed that the OFC appears to contain information about the model of the task structure and influence the dopamine neurons so that they can compute reward prediction error according to the model.

To simulate this new task involving action selection, we need to extend our circuit model in three directions: (1) we need to consider how the dopamine neurons in the VTA, instead of those in the SNc, compute reward prediction error, because the abovementioned studies (Roesch et al., 2007; Takahashi et al., 2011) mainly examined the VTA cells; (2) we should incorporate the corticobasal ganglia–thalamocortical feedback loop instead of the pathway to the SC to explain nose/arm/body movements rather than eye movements; and (3) we should incorporate a mechanism for action selection. As for the first point, here we propose that the same corticostriatal mechanism for the computation of the TD of value predictions, which we proposed originally for the SNc dopamine neurons (Morita et al., 2012) (Fig. 7*A*), operates also for the VTA dopamine neurons, with different intermediate nuclei as shown in Figure 7*B*. This is based on the *in vivo* findings that neurons in the GPb (Hong and Hikosaka, 2008), LHb (Matsumoto and Hikosaka, 2007, 2009), and rostromedial tegmental nucleus (RMTg)/tail of the VTA (tVTA) (Jhou et al., 2009b; Hong et al., 2011), as well as GABAergic neurons in the VTA (Cohen et al., 2012), appear to represent negative reward signal or sign-reversed reward prediction error signal, as well as on the anatomical and/or physiological demonstrations that there are excitatory connections from the GPb to the LHb (Hong and Hikosaka, 2008; Shabel et al., 2012) and from the LHb to the RMTg/tVTA (Jhou et al., 2009a; Hong et al., 2011; Lammel et al., 2012) and inhibitory connections from the RMTg/tVTA to the VTA/SNc dopamine neurons (Jhou et al., 2009a; Hong et al., 2011); transient inhibitory influence of the dopamine neurons in the SNc and the VTA on the LHb neuronal firing has also been reported (Shen et al., 2012). Although the upstream of the GPb is currently unknown, it is conceivable that it receives inputs from both the direct and indirect pathways of the basal ganglia, given that the GPb is at the borders of the internal segment of GP (GPi) (Hong and Hikosaka, 2008), which receives inputs from both of these pathways. Notably, in the abovementioned study (Takahashi et al., 2011), stimulation of the OFC caused either excitatory or inhibitory effects on the firing of the VTA dopamine neurons, with the latter more frequent, and the inhibitory effects typically lasted for a few to several hundred milliseconds after the offset of stimulation. These results seem to be consistent with our model (Fig. 7*B*): the sustained inhibition is considered to reflect the (indirect) influence of the CPn/PT cells, which are presumably able to sustain their activity via strong recurrent excitation in our model (Morita et al., 2012).

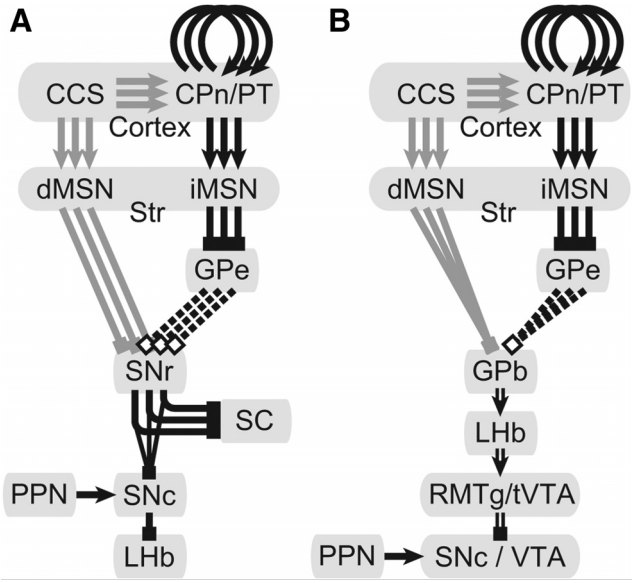

Figure 7. Putative pathways conveying positive and negative reward signals, with the common corticostriatal mechanism for computing the TD of reward predictions. ***A***, Pathway that conveys positive reward signals and computes reward prediction error in the SNc (Morita et al., 2012), which is then transmitted to the LHb with a sign reversal. The arrows, filled squares, and open squares indicate excitation, inhibition, and disinhibition, respectively. ***B***, Pathway that conveys negative reward signals, which are transmitted from the GPb to the LHb, the RMTg (also referred to as the tVTA; Bourdy and Barrot, 2012), and then to the SNc and the VTA with a sign reversal. The model used for the simulations in the present study does not explicitly describe structures between the striatum (Str) and the SNc/VTA (see Materials and Methods), and it could thus simultaneously represent both of these two pathways, except for suggested mutually inhibitory interactions between the SNc and the LHb.

Next, regarding the second and the third points (corticobasal ganglia–thalamocortical feedback loop and a mechanism for action selection), we proposed (Morita et al., 2012) that the inputs from this feedback loop specifically target the (apical tuft dendrites of the) CPn/PT cells based on anatomical and morphological findings (explained in detail by Morita et al., 2012). Given this circuit architecture, here we propose that either of the following mechanisms for action selection (or more precisely, action plan selection; see below) can be implemented (Fig. 8*A–D*) as follows.

(1) A plan of action to be executed, $A(t_i)$, is selected outside of the circuit that we consider, according to the predicted values of possible action plans in a soft-max manner (Fig. 9*A*, right graph), and only that selected action plan is loaded onto a subset of CCS cells (Fig. 8*A*). The selected action plan $A(t_i)$ is then sent for execution through boosting of the corresponding subset of CPn/PT cells by the

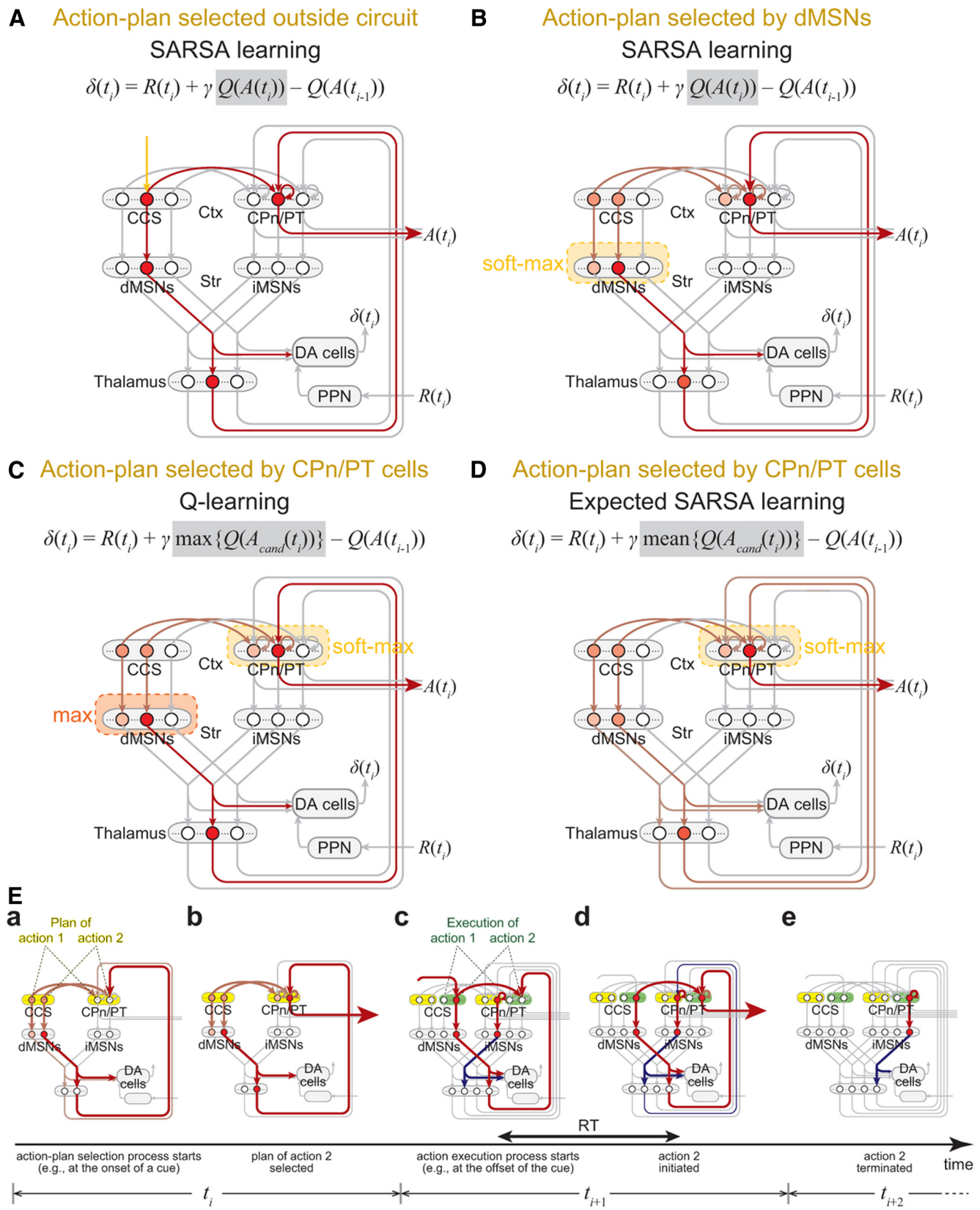

Figure 8. Reward prediction error computation in parallel with action selection and/or execution in the corticobasal ganglia circuit. ***A–D***, Four different (extreme) possibilities are described. ***A***, An action plan selected at the outside of the circuit considered here is loaded onto a subset of CCS cells and then sent for execution through the direct pathway–thalamic loop and the CCS ⟶ CPn/PT connections. In the meantime, the dopamine neurons compute $R(t_i) + \gamma Q(A(t_i)) - Q(A(t_{i-1}))$, which corresponds with reward prediction error defined in the SARSA algorithm (under the condition in which the input– output functions f_1 and f_2 of MSNs operate in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists; the same applied to all the cases). ***B***, Candidates of action plans are loaded onto different subsets of CCS cells, and an action plan is selected in the CCS– dMSNs pathway, through feedforward inhibition (and possibly also lateral and feedback inhibition), with the probability depending on its predicted value, implementing a soft-max operation. The selected action plan is then sent for execution (*Figure legend continues*.)

direct CCS → CPn/PT connections and the basal ganglia–thalamocortical feedback, which is released from inhibition by the inputs from the CCS cells to the dMSNs/direct pathway.

(2) Candidates of action plans are loaded onto different subsets of CCS cells (Fig. 8*B*). Then an action plan is selected in the CCS–dMSNs pathway, through feedforward inhibition (Parthasarathy and Graybiel, 1997; Gittis et al., 2010) (and possibly also lateral and feedback inhibition), with the probability depending on its predicted value, implementing a soft-max operation. The selected action plan will eventually be sent for execution through the basal ganglia–thalamocortical loop and the CPn/PT cells (without perturbation at the stage of CPn/PT cells).

(3) Candidates of action plans are loaded onto different subsets of CCS cells, as in case 2. However, different from case 2, the CCS–dMSNs pathway implements the (hard) max operation rather than a soft-max operation (Fig. 8*C*), and thus an action plan that currently has the maximum value is selected at the stage of dMSNs after a brief initial transient phase. Meanwhile, receiving the initial dMSN-direct pathway–thalamic inputs that represent predicted values of each action plan (Fig. 8*Ea*), CPn/PT cells begin to compete with each other, and an action plan that will eventually be sent for execution is selected in a soft-max manner; this selected (to-be-executed) action plan is not necessarily the same as the one selected by dMSNs (i.e., the one with the maximum value), because the two selection processes presumably proceed rather separately [i.e., if a CPn/PT population corresponding to an action plan different from the one with the maximum value survives the winner-take-all competition through recurrent attractor dynamics (cf. Wang, 2002), the winner will not be changed by the dMSN–trans-thalamic inputs].

(4) The CCS–dMSNs pathway does not implement any computation, neither a soft-max nor the max operation (Fig. 8*D*). A soft-max operation is implemented through competition among the CPn/PT cells.

Along with action (plan) selection, TD error for action (plan) values (Sutton and Barto, 1998) is computed and represented in the dopamine neurons. Importantly, different forms of reward prediction error are computed in the different cases raised above, and they correspond with different algorithms of reinforcement learning. In cases 1 and 2, the computed signal is $R(t_i) + \gamma Q(A(t_i)) - Q(A(t_{i-1}))$, where $R(t_i)$ is reward obtained at time t_i, and $Q(A(t_i))$ and $Q(A(t_{i-1}))$ are predicted values of $A(t_i)$ and $A(t_{i-1})$, respectively (under the condition in which the input–output functions f_1 and f_2 of the MSNs operate in their suprathreshold linear regimens and are not affected by dopamine receptor antagonists; the same is applied to all the cases). This signal corresponds with the TD error for the SARSA algorithm. In contrast, in case 3, the computed signal is $R(t_i) + \gamma \max\{Q(A_{\text{cand}}(t_i))\} - Q(A(t_{i-1}))$, where $A_{\text{cand}}(t_i)$ are candidates of action plans, and this signal corresponds with the TD error for *Q*-learning [to be precise, this formula is true only for a portion of the time step t_i after the completion of the max operation (Fig. 8*Eb*), but in the model equations, we just assumed this formula as a dopamine neuronal response at t_i as an approximation]. In case 4, the computed signal is $R(t_i) + \gamma \text{ mean}\{Q(A_{\text{cand}}(t_i))\} - Q(A(t_{i-1}))$, which corresponds with the TD error for the expected SARSA learning (van Seijen et al., 2009) (or "sum" instead of "mean"; but these two could become equivalent by changing γ). Which of these cases actually operates would depend on the fine properties of circuits and neurons, which may differ in different portions of the corticobasal ganglia loop, and/or on the conditions of neuromodulation. Also, the above four are all extreme cases, and actual operation in the brain can be somewhat between them; for example, we assumed that there is no perturbation at the stage of the CPn/PT cells in case 2, but in reality, there is likely to be at least some perturbation. Notably, soft-max operation at the stage of the CPn/PT cells could be achieved by competitive dynamics of a large number of cortical neurons (cf. Wang, 2002). It could also be related to suggested variability generation in the lateral magnocellular nucleus of the nidopallium in songbirds (Olveczky et al., 2005), which could be homologous to the mammalian CPn/PT cells. In any case, exploring ways of implementations of the abovementioned possibilities could be a nice theme for biophysical modeling study in the future, but here we focus on case 3 in which the circuit implements *Q*-learning, which has been suggested in the original study (Roesch et al., 2007) as mentioned above.

Importantly, there can be a certain time interval between action (plan) selection (in a brain) and action execution, as reproduced by typically using delayed response tasks in the laboratory. Therefore, action (plan) selection and action execution may not always be a single inseparable process but rather constitute serially operating (in an approximate sense) processes. Specifically, cortical neurons that represent plan (preparation) for a particular motor action would become selectively activated first, and thereafter (after a scheduled delay if there is) those "action-plan (action-preparation)" neurons will activate a different subpopulation of cortical neurons that represent execution of that partic-

←

(*Figure legend continued.*) without perturbation at the stage of CPn/PT cells, while the dopamine neurons compute the same SARSA error signal as the case of ***A***. ***C***, Candidates of action plans are loaded onto different subsets of CCS cells in the same manner as in ***B***, but the CCS–dMSNs pathway implements the (hard) max operation rather than a soft-max operation assumed in ***B***. Consequently, an action plan that currently has the maximum value is selected at the stage of dMSNs after a brief initial transient phase, and thereby the dopamine neurons compute $R(t_i) + \gamma \max\{Q(A_{\text{cand}}(t_i))\} - Q(A(t_{i-1}))$, which corresponds with reward prediction error defined in *Q*-learning. Meanwhile, receiving the initial dMSN– direct pathway–thalamic inputs that represent predicted values of each action plan (see ***Ea***), CPn/PT cells begin to compete with each other, and a plan of the action that will eventually be executed is selected in a soft-max manner; this selected (to-be-executed) action plan is not necessarily the same as the one selected by dMSNs (i.e., the one with the maximum value), because the two selection processes presumably proceed rather separately. ***D***, Candidates of action plans are again loaded onto different subsets of CCS cells, but the CCS– dMSN pathway implements neither a soft-max nor the max operation. Instead, values of the action plan candidates are computed in the dMSNs and sent to the dopamine neurons as they are. The dopamine neurons therefore compute $R(t_i) + \gamma \text{ mean}\{Q(A_{\text{cand}}(t_i))\} - Q(A(t_{i-1}))$, which corresponds with the TD error for the expected SARSA learning (van Seijen et al., 2009) (or "sum" instead of "mean"; but these two could become equivalent by changing γ). Receiving the thalamic feedback inputs, the CPn/PT cells compete to select an action plan that is to be executed, implementing a soft-max operation. ***E***, Schematic illustration of the presumed operation of the circuit, with a consideration that action plan selection and action execution are serially operating separable (in an approximate sense) processes (***a*** → ***b***). Among two action-plan candidates ("plan of action 1" and "plan of action 2"), the plan of action 2 is selected through the presumed *Q*-learning-compatible mechanism, and the corresponding subset of CPn/PT cells becomes selectively activated. These activated CPn/PT cells representing the plan of action 2 then activate a subset of CCS cells that represent execution of action 2, via presumably hard-wired intracortical (possibly inter-areal) connections (***c***). Subsequently, those CCS cells activate CPn/PT cells that also represents execution of action 2, via both the direct intracortical connections and the CCS → dMSNs → SNr → thalamus → CPn/PT loop pathway, and thereby action 2 is initiated (***d***). Those CPn/PT cells will keep their activity via strong recurrent excitation and send inputs to iMSNs so as to provide the dopamine neurons with a negative-signed signal of the value of the executed action via the indirect pathway (***e***). Notably, reaction time corresponds to the duration between ***c*** and ***d*** and is thus modulated by the activity of dMSNs that receive inputs from the CCS cells representing execution of action 2 (***c***). Ctx, Cortex; Str, striatum.

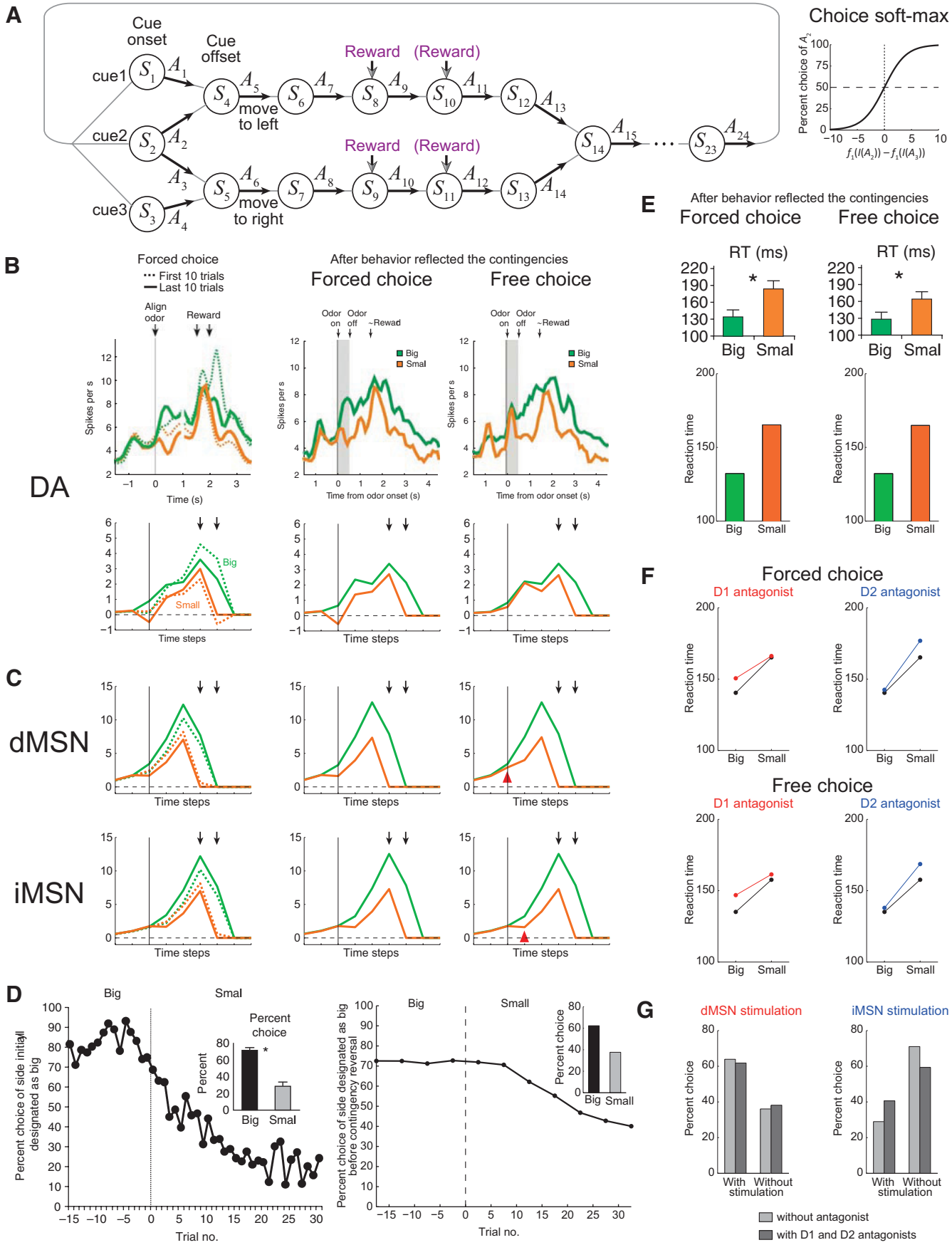

Figure 9. Simulation of a task involving action selection considered by Roesch et al. (2007) and Takahashi et al. (2011): explanation of the neural and behavioral data and the predictions. ***A***, Left, Presumed state transitions during individual task trials (similar to the one proposed in a model developed by Takahashi et al., 2011). S_1, S_2, . . . represent subjects' states (*Figure legend continues.*)

ular action, i.e., "action-execution" neurons, via presumably hard-wired intracortical (possibly inter-areal) connections (e.g., neurons that represent plan/preparation for moving to the left will activate neurons that represent actual movement to the left). Figure 8*E* shows how this scheme goes well with our model, in the case of *Q*-learning (Fig. 8*Ea*,*b* is an extended version of Fig. 8*C* with a higher temporal resolution). As shown in Figure 8*E*, a subset of CPn/PT cells that represents plan/preparation for a particular motor action ("action 2") becomes selectively activated (Fig. 8*Ea* → *Eb*), through the presumed mechanism for *Q*-learning-compatible action (plan) selection in the corticobasal ganglia circuit (Fig. 8*C*), and those CPn/PT cells then activate a subset of CCS cells representing execution of that particular action (Fig. 8*Ec*), which can be located in a different cortical region/area. Indeed, a recent study has shown that a portion of CPn cells contribute preferentially to intracortical connections from higher to lower motor cortical areas (Ueta et al., 2013): we assume that such inter-areal connections drive CCS cells in the recipient area, either directly or indirectly via upper layer neurons. The activated CCS cells will then drive CPn/PT cells that also represent execution of that action, via both the direct intracortical connections and the CCS → dMSNs → SNr → thalamus → CPn/PT loop pathway, and thereby the action is initiated (Fig. 8*Ed*). Those CPn/PT cells will keep their activity via strong recurrent excitation and send inputs to iMSNs so as to provide the dopamine neurons with a negative-signed signal of the value of the executed action via the indirect pathway (Fig. 8*Ee*).

Notably, our model equation describes activity of dMSNs, iMSNs, and the dopamine neurons at discrete time points (t_i, t_{i+1}, and t_{i+2} in Fig. 8*E*), but this is certainly a very rough approximation, and there are many issues regarding detailed biophysical processes that need to be clarified in the future. In particular, it is not trivial how the CPn/PT cells representing action execution initially send powerful inputs to the pyramidal tract so that action can be initiated (Fig. 8*Ed*) but then subsequently keep sending major inputs to iMSNs, but not to the pyramidal tract, after the termination of the action (Fig. 8*Ee*). One possibility is that the temporal pattern of CPn/PT spikes plays a critical role. Specifically, at the early phase (Fig. 8*Ed*), the CPn/PT cells presumably receive both inputs directly from CCS cells and those via the trans-thalamic projections, which preferentially target layer 1 (Kuramoto et al., 2009) in which distal apical dendrites of the CPn/PT cells are located. Such convergence of basal and distal apical inputs can cause burst firing of pyramidal cells (Larkum et al., 1999), and burst can be efficiently transmitted down the pyramidal tract through corticospinal synapses given that, at those synapses, postsynaptic response is known to be facilitated when spikes arrive with several milliseconds intervals (Meng et al., 2004), which are typical for intraburst spikes. In contrast, at the later phase (Fig. 8*Ee*), activity of the CPn/PT cells is sustained mainly through recurrent excitation. They would also receive recurrent inhibitory inputs onto cell domains including distal apical dendrites (Silberberg and Markram, 2007). Such dendritic inhibition could be oscillatory as a population, and combined with hyperpolarization-activated current (I_h) that is especially rich in corticospinal (CPn/PT-type) pyramidal cells (Sheets et al., 2011), could potentially disturb burst generation and cause oscillatory spiking at ~15–25 Hz (beta range) and thereby drastically change efficacies of different downstream pathways, as suggested by a recent modeling study (Li et al., 2013). This burst/beta-dependent switching of information flow seems to be in line with the well-known fact that corticospinal–muscular beta rhythm is absent at the time of movement but appears prominently after the movement (Baker, 2007), as well as the exaggerated beta activity in the basal ganglia in Parkinson's disease (Brown, 2007). Switching of information flow at the previous time point (Fig. 8*Eb* → *Ec*) might also depends on a similar mechanism, in which burst-dependent induction of local spikes in pyramidal dendrites (Polsky et al., 2009) may achieve burst detection. These details are expected to be explored in the future (see also the additional discussion about the discrete time representation in Materials and Methods, at the end of Elaborated model for the saccade task).

←

(*Figure legend continued.*) that are defined by external events (i.e., cue onset, cue offset, reward bolus), their own movements, or internally (i.e., internal state), and $A_1, A_2, \ldots$ represent option(s) that can be taken at each state (e.g., "plan/prepare for moving to the left," "move to the left," or "keep rest"). At the beginning of a trial (at the leftmost in the diagram), the subject is presented with one of three odor cues, two of which (cues 1 and 3) indicated that reward will be given if the animal entered either the left-side or right-side well, respectively (forced-choice trial), whereas the remaining cue (cue 2) indicated that the animal will be rewarded in both of the wells (free-choice trial). In any case, the amount of reward, either small (1 bolus of sucrose solution at S_8 or S_9) or large (2 boluses, with the second bolus given successively at S_{10} or S_{11}), was determined according to the predetermined direction-amount contingency that was fixed during a block (at least 60 trials in the experiments; 120 trials in our simulations) and then reversed. Forced-choice trials and free-choice trials were pseudorandomly intermingled. In our model, when the subject enters each state, subset(s) of CCS cells in the OFC are assumed to represent combination(s) of that state and options that can be taken there [e.g., at state S_2, a subset represents "$S_2 - A_2$" (plan/prepare for moving to the left) and another represents "$S_2 - A_3$" (plan/prepare for moving to the right)]. Subsequently, the circuit is assumed to operate as we considered in Figure 8*C*, and one of the options is selected (and executed if it is a movement); if there is just a single option, it is selected (and executed). In the meantime, the dopamine neurons compute TD reward prediction error (regardless of whether there are multiple options or a single option), according to which the strengths of corticostriatal (i.e., CCS– dMSNs and CPn/PT–iMSNs) connections are plastically modified. Notably, the three time points for S_2 (cue onset), S_4/S_5 (cue offset), and S_6/S_7 correspond to t_i, t_{i+1}, and t_{i+2} in Figure 8*E*, and thus reaction time is presumably modulated by the activity of dMSNs at cue offset. Right, Assumed soft-max function for choice. ***B***, Dopamine neuronal response sorted by different experimental conditions: the top and bottom rows show the results of the experiments (Roesch et al., 2007) and our simulations, respectively. The vertical black line indicates the timing of cue onset. Left, Dashed and solid lines indicate the average response of dopamine neurons in the first and last 10 forced-choice select-large (green) or select-small (orange) trials in the blocks, respectively. Middle and Right, Green and orange lines indicate the average response of dopamine neurons in forced-choice (middle) or free-choice (right) select-large and select-small trials, respectively. To control for learning and also for the possibility that disadvantageous choices might be more often early in the block, only trials after the ratio of selecting more valuable option exceeds 50% were included, and each free-choice trial was paired with the immediately preceding and following forced-choice trial of the same reward amount [as done in the original work (Roesch et al., 2007)]. ***C***, Population activity of dMSNs (top row) and iMSNs (bottom row) predicted from our model, sorted by the same experimental conditions as considered in ***B***. ***D***, The impact of reward amount on choice behavior in free-choice trials. Line graphs show the ratio (percentage) of choices before and after the switch from big reward to small reward (indicated by the vertical line), and the inset bar graphs show the percentage choice for large versus small reward across all free-choice trials. The left and right panels show the results of the experiments (Roesch et al., 2007) and our simulations, respectively. ***E***, Average reaction time in the forced-choice (left column) and free-choice (right column) trials included in the data shown in the middle and right panels of ***B***, respectively. Top row shows the experimental results (Roesch et al., 2007), and the bottom row shows our simulation results. ***F***, Prediction from our model about the effects of D_1 (left) or D_2 (right) antagonist on the reaction time in forced-choice trials (top) and free-choice trials (bottom). ***G***, Prediction from our model about the effects of optogenetic stimulation of dMSNs (left) or iMSNs (right) on choice behavior. In a new set of simulations, in one of the two wells (left and right), virtual optogenetic stimulation was applied to either dMSNs or iMSNs coincidently with reward (at S_8 or S_9), in addition to giving an extra bolus of reward at the subsequent timing in both of the wells; the contingency between the stimulation on/off and the location of the well was fixed for a block and alternated across blocks. The bar graphs show the ratio (percentage) of choices for with versus without optogenetic stimulation across all free-choice trials, either without dopamine receptor antagonist (light gray bars) or with both D_1 and D_2 antagonists (dark gray bars). The top rows in ***B*** and ***E*** and the left panels of ***D*** were taken from Roesch et al. (2007).

Simulation of a task involving action selection: reproductions and predictions

In the experiments (Roesch et al., 2007; Takahashi et al., 2011), at the beginning of each task trial, the subject should make a nose poke to be presented with one of three odor cues, two of which indicated that reward will be given if the animal entered either the left or right well, respectively (forced-choice trial), whereas the remaining cue indicated that the animal will be rewarded in both of the directions (free-choice trial). In any case, the amount of reward, either small (one bolus of sucrose solution) or large (two boluses given sequentially), was determined according to the predetermined direction-amount contingency that was fixed during a block (at least 60 trials in the experiments; 120 trials in our simulations) and then reversed. Forced-choice trials and free-choice trials were pseudorandomly intermingled. Takahashi et al. (2011) examined the effects of a lesion of the OFC on the dopamine neuronal activity and subjects' behavior. Comparing the results with a computational model developed within the work and also combining electrical stimulation, they revealed that, in the intact subjects, the OFC contains information about the model of the task structure, i.e., a diagram for transitions between different states, and influences the VTA dopamine neurons so that they can compute reward prediction error according to the model.

According to these results (Takahashi et al., 2011), we assumed that a state-transition diagram as shown in Figure 9*A*, which is similar to the one considered by Takahashi et al. (2011), is represented in the OFC. In this diagram, S_1, S_2, . . . represent the states of the subject that are defined by external events (i.e., cue onset, cue offset, reward bolus), their own movements, or internally (i.e., internal state), and A_1, A_2, . . . represent option(s) that can be taken at each state (e.g., "plan/prepare for moving to the left," "move to the left," or "keep rest"). Then, according to our proposal on the architecture of corticobasal ganglia circuits (Morita et al., 2012) and a demonstration of the existence of CCS cells and CPn/PT cells in the OFC (Hirai et al., 2012), we assumed that the CCS cells and CPn/PT cells in the OFC influence the VTA dopamine neurons through the pathways shown in Figure 7*B*, and the circuit operates in the following manner. Entering each state, subset(s) of CCS cells represents combination(s) of that state and option(s) that can be taken there [e.g., a subset represents "$S_2 - A_2$" (plan/prepare for moving to the left) and another represents "$S_2 - A_3$" (plan/prepare for moving to the right)]. Subsequently, according to the mechanism that we considered in the above (Fig. 8*C*), one of the options is selected (and executed if it is a movement); if there is just a single option, it is selected (and executed). In the meantime, the dopamine neurons compute TD reward prediction error (regardless of whether there are multiple options or a single option), according to which the strengths of corticostriatal (i.e., CCS– dMSNs and CPn/PT–iMSNs) connections are plastically modified.

Making some additional assumptions including time-dependent decay of the changes in the strength of corticostriatal connections (for details, see Materials and Methods), we found conditions with which the model can well reproduce important features of the dopamine neuronal activity reported in the work that we have modeled (Roesch et al., 2007), including relative magnitude depending on conditions (forced-choice vs free-choice and select-big vs select-small) at the timings of cue and reward and also more overall within-trial time course of the activity (Fig. 9*B*). In particular, the model well reproduced the experimental observation indicative of *Q*-learning (Roesch et al., 2007) that the dopamine neurons show differential responses to the big- and small-reward-predicting cues in forced-choice trials, whereas the response to the free-choice cue is similar to the response to the big-reward-predicting forced-choice cue regardless of the obtained reward amount. With the same parameters that were used to reproduce these neural data, our circuit model also successfully reproduced the observed shorter reaction time in select-big trials than in select-small trials in both forced-choice and free-choice trials (Fig. 9*E*), if we assume that reaction time is determined by the activity of dMSNs at the timing of cue offset, in a manner similar to our models for the saccade task described above. Moreover, reinforcement of a choice leading to big reward has also successfully occurred (Fig. 9*D*) [as shown in this figure, the choice performance of the model is somewhat worse than the animals' performance in the study by Roesch et al. (2007), but it is more similar to the performance in the study by Takahashi et al. (2011) using the same task].

Notably, however, the success in explaining the neural and behavioral data itself is not unique to the specific circuit architecture of our model, and, in fact, it has already been done in the model developed by Takahashi et al. (2011). Instead, what is new to our model is that it can predict the activity of dMSNs and iMSNs, as well as the effects of specific manipulations of either of these cell populations. As shown in Figure 9*C*, it is predicted that activation of dMSNs precedes that of iMSNs, reflecting the unidirectional connectivity from the CCS cells to the CPn/PT cells (Morishima and Kawaguchi, 2006) (the data shown in Fig. 9*C* should be regarded as the population activity; for details, see Materials and Methods). In free-choice trials, the activity of dMSNs initially increases in a similar rate to the case of forced-choice big-reward trials regardless of whether big-reward option will be subsequently selected or not (Fig. 9*C*, top right, red arrowhead). In contrast, it is not the case for the activity of iMSNs (Fig. 9*C*, bottom right, red arrowhead). Such a difference reflects the assumption about how *Q*-learning is implemented in the corticobasal ganglia circuit (Fig. 8*C*). Next, we examined effects of dopamine receptor antagonists in this model. Assuming the same effects of D_1 and D_2 antagonists on the responsiveness of dMSNs and iMSNs as assumed in our models for the saccade task (Fig. 3), simulations revealed that D_1 antagonist prominently slows a movement leading to big reward but has little effect on the one leading to small reward (Fig. 9*F*, left), whereas the opposite is the case for D_2 antagonist (Fig. 9*F*, right) in both forced-choice trials (Fig. 9*F*, top row) and free-choice trials (bottom row). It is thus predicted that the distinct effects of dopamine receptor antagonists on the reaction time observed in the forced-choice saccade task (Nakamura and Hikosaka, 2006) can also occur in the task involving action selection and also with different effectors.

We further conducted virtual optogenetic stimulation of dMSNs or iMSNs using the same model, trying to qualitatively explain the results of another recent study (Kravitz et al., 2012), which selectively stimulated either dMSNs or iMSNs when optogenetically tagged mice touched a trigger and found that mice learned to seek and avoid self-stimulation of dMSNs and iMSNs, respectively, even under the presence of D_1 and D_2 antagonists. We assumed that, in one of the two wells (left or right), stimulation of either dMSNs or iMSNs was applied in coincidence with a reward bolus, in addition to a subsequent extra bolus of reward in both of the wells [i.e., except for the MSN stimulation, both of the wells are in the "big reward" condition in the original experiments (Roesch et al., 2007; Takahashi et al., 2011)]; the contingency between the location of the well and the stimulation of MSNs was fixed during a block and alternated across blocks. Through simulations, it was shown that a well associated with

dMSN stimulation becomes more frequently chosen, whereas a well associated with iMSN stimulation becomes more frequently avoided, even under the presence of D_1 and D_2 antagonists (Fig. 9*G*), successfully explaining the essential findings of the optogenetic stimulation study (Kravitz et al., 2012). Notably, in our simulations for conditions with the antagonists, if the virtual optogenetic stimulation is applied on the "small reward" condition (i.e., when the subsequent extra bolus of reward is omitted), dMSN stimulation still causes attraction but iMSN stimulation fails to cause aversion (data not shown). This is presumably due to the threshold of the dMSNs' input-output function assumed in the model: if the corticostriatal connections are not sufficiently strengthened, corticostriatal inputs to dMSNs remain to be subthreshold and cannot cause any choice bias. This seems to be a model-dependent phenomenon (i.e., changing the threshold might change the results), but could still have an implication: when subject expects little reward in the near future, s/he might not be able to learn advantageous (i.e., not so good, but not very bad) choice strategy well (see Potjans et al., 2011 for a discussion on this issue in relation to the asymmetry of positive and negative dopamine response). There can be certain mechanisms to upregulate the activity of dMSNs to supra-threshold level, in particular, inputs from cortical neurons showing task-related activity and/or increase of the striatal tonic dopamine (cf. Niv et al., 2007), so that subject can still learn advantageous choice strategy even in such situations.

Discussion

Dopamine has been suggested to be crucially involved in the control of motivation and reinforcement learning, but how the activity of dopamine neurons itself is controlled has remained elusive. We tried to resolve this issue by constructing a closed-circuit model of the corticobasal ganglia system based on recent findings on the intracortical and corticostriatal circuit architectures. Through numerical simulations, we showed that our model successfully reproduces the observed across- and within-trial changes in the dopamine neuronal response that represents reward prediction error, as well as changes in reaction time depending on expected reward amount and changes in choice depending on learning of reward values. Moreover, our model could also explain the observed distinct effects of manipulations of the direct or indirect pathway striatal neurons, on reaction times as well as on choices. Importantly, the results of this study challenges a current popular view on the functions of the distinct pathways of the basal ganglia as we explain below.

Roles of the direct and indirect pathways of the basal ganglia

A current popular hypothesis regarding the function of the basal ganglia is that the direct and indirect pathways are involved in appetitive ("go") and aversive ("no-go") learning, respectively (Frank et al., 2004; Frank, 2005). The experimental result (Kravitz et al., 2012) that mice learned to seek and avoid stimulation of their own dMSNs and iMSNs, respectively, was interpreted to support this hypothesis, given an assumption that dopamine signaling was "bypassed" by the optogenetic stimulation of MSNs (Paton and Louie, 2012). Our model provides an alternative view (Table 1). According to our model, optogenetic stimulation would cause significant phasic response of dopamine neurons in specific ways: stimulating dMSNs or iMSNs is expected to cause positive or negative phasic response of dopamine neurons, respectively, as if the touch that the animal has just made has led to positive or negative reward prediction error. This could explain the observed seeking and avoidance of dMSN and iMSN self-stimulation, respectively (Kravitz et al., 2012), as we qualitatively showed in the above (Fig. 9*G*). Our hypothesis potentially could also explain several other recent empirical findings that have so far been explained by the direct-go/indirect-no-go hypothesis.

Table 1. Functions of the direct and indirect pathways of the basal ganglia: a popular view versus a new view suggested from our model

	Direct pathway	Indirect pathway
Popular view		
Operate for	Upcoming action	Upcoming action
How to operate	Execute ("go")	Inhibit ("no go")
If stimulated	Action execution/facilitation	Action inhibition
Our model		
Operate for	Upcoming action	Ongoing/executed action
How to operate	Execute ("go")	Terminate
If stimulated	Action execution/facilitation and positive reward prediction error	Action termination and negative reward prediction error

A key feature that distinguishes our hypothesis from the go/no-go explanation is the functional role of the indirect pathway: in our view, the indirect pathway should represent the predicted value of the state/action at a previous time point. This time delay comes from the presumed sustained firing of the upstream CPn/PT cells (Morita et al., 2012), and it is expected to manifest as delayed firing of iMSNs compared with dMSNs (Fig. 9*C*), although at a coarser timescale, dMSNs and iMSNs will presumably show phasic activation at rather similar timings (around target/cue presentation; Fig. 6*D*,*E*). Notably, our view is broadly consistent with previous suggestions that the CPn/PT → indirect pathway may receive an efference copy signal and might serve action termination (Lei et al., 2004; Graybiel, 2005; Reiner et al., 2010).

Another (although closely related) important difference between the two hypotheses is the way of dopamine dependence of corticostriatal synaptic plasticity. Whereas the go/no-go hypothesis assumes that phasic dopamine response induces plasticity of the opposite directions for dMSNs and iMSNs (i.e., increase of dopamine strengthens and weakens the synapses on dMSNs and iMSNs, respectively), we assume that the induced plasticity should be the same direction (increase of dopamine strengthens both dMSN and iMSN synapses; for the potential validity of this assumption, see Materials and Methods). Notably, our hypothesis incorporates the difference in the cortical input sources between dMSNs and iMSNs, namely, inputs coming preferentially from CCS cells and from CPn/PT cells, respectively (Lei et al., 2004; Reiner et al., 2010). The go/no-go hypothesis does not take this into account. It is true that a lot of differences between the plasticity on dMSNs and iMSNs have been demonstrated (Gerfen and Surmeier, 2011), and to be fair, plasticity induction in these two cell types appears to be toward the opposite directions, in line with the go/no-go hypothesis, at least under certain experimental conditions (Shen et al., 2008). However, here we propose that these differences are to implement the same direction of dopamine-dependent plasticity for the synapses receiving different temporal patterns of cortical inputs (CCS–dMSN: early/transient; CPn/PT–iMSNs: late/sustained) rather than to implement the opposite directions of dopamine dependence. Future experiments are expected to test these alternative hypotheses.

Roles of dopamine in motivational control

Regarding the dual role of dopamine in motivational control and reinforcement learning, an intriguing idea has been proposed

(Niv et al., 2007) that the tonic level of dopamine represents response vigor, a manifestation of motivation, whereas its phasic release represents reward prediction error. In our model, reaction time is assumed to be quasi-inversely related to the stimulus-induced response of dMSNs in the presence of tonic dopamine, and D_1 antagonist induces an increase of reaction time in large-reward blocks (Fig. 4*D*) directly through its effect on the input–output function of dMSNs under tonic dopamine, in line with the idea proposed by Niv et al. (2007). Conversely, the effect of D_2 antagonist appears indirectly, through the change in the activity of iMSNs and their downstream dopamine neurons as explained previously. Notably, such distinct ways of operation of D_1 and D_2 antagonists in the model lead to earlier manifestation of the effect of D_1 antagonist than that of D_2 antagonist (approximately three trials after block switch in Fig. 4*D*; approximately five trials in Fig. 5*D*), which appears to be consistent with the experimental observations (Figs. 4*E*, 5*E*).

Motivation is a complex entity, and there should be many aspects that cannot be captured by reaction time or response vigor. As mentioned previously, according to our model, the degree of reward time discount is determined by the balance between the direct and indirect pathways of the basal ganglia (Fig. 6*C*), which can be regulated by striatal tonic dopamine (Albin et al., 1989; DeLong, 1990). Increase in the dopamine level would lead to a milder discount of future rewards, which could be subjectively felt as an enhancement of motivation. This is just an example, and, in fact, how dopamine is involved in a multitude of motivational processes is a challenging future issue. Given the closed-circuit architecture, our model can hopefully serve as an useful test bed in exploring physiological mechanisms of not only extrinsically induced motivation but also how motivation is intrinsically generated and how it interacts with extrinsic motivation (Murayama et al., 2010).

Roles of dopamine in reinforcement learning and decision making

Phasic dopamine response appears to represent reward prediction error and is thought to play crucial roles in reinforcement learning and value-based decision making (Glimcher, 2011). In the meantime, several areas in the frontal cortex, including the OFC, have been suggested to be crucially involved in reward-guided learning and decision making (Rushworth et al., 2011). A recent study (Takahashi et al., 2011) has elucidated an exact relationship between the dopamine neurons and the OFC. Specifically, it has been suggested (Takahashi et al., 2011) that information about task structure, i.e., the model of state transitions, is represented in the OFC, and it influences the VTA dopamine neurons so that they can compute reward prediction error according to the model, presumably via intermediate structures such as the ventral striatum in which the value of the states would be computed. Following this suggestion, we proposed a specific circuit mechanism at the resolution of intermingled neural subpopulations within each structure, providing rich predictions that are expected to be tested in future experiments. How the model of state transitions itself is acquired is currently unknown and is also expected to be explored in the future.

Our present model consists of only a single corticobasal ganglia–midbrain loop circuit. However, in reality, different parts of the loop have been suggested to be functionally specialized. Indeed, it has been suggested previously (Roesch et al., 2007) that, although the dopamine neurons in the VTA appear to encode TD error for *Q*-learning (Roesch et al., 2007), those in the SNc were shown to represent the error signal for SARSA (Morris et al., 2006). Moreover, some populations of dopamine neurons may not represent reward prediction error (Bromberg-Martin et al., 2010). Likewise, different parts of the striatum may be specialized for state versus action learning (O'Doherty et al., 2004) and/or for different timescales (Tanaka et al., 2004), and an even further functional specialization would exist in the cortex (Rushworth et al., 2011). With regard to the tasks that we modeled, in addition to the model of the presumed state transitions, knowledge that if one action/location is associated with small reward, the other should lead to large reward (and vice versa) may also be acquired and represented in certain cortical places (Hong and Hikosaka, 2011) so that a subject can change behavior by experiencing only a single trial after a reversal of the location–reward contingency, as observed in experiments after extensive training (Watanabe and Hikosaka, 2005). How it is implemented and how it interacts with the mechanisms that we proposed remain as important future issues.

References

Aggarwal M, Hyland BI, Wickens JR (2012) Neural control of dopamine neurotransmission: implications for reinforcement learning. Eur J Neurosci 35:1115–1123. CrossRef Medline

Albin RL, Young AB, Penney JB (1989) The functional anatomy of basal ganglia disorders. Trends Neurosci 12:366–375. CrossRef Medline

Azdad K, Gall D, Woods AS, Ledent C, Ferré S, Schiffmann SN (2009) Dopamine D2 and adenosine A2A receptors regulate NMDA-mediated excitation in accumbens neurons through A2A-D2 receptor heteromerization. Neuropsychopharmacology 34:972–986. CrossRef Medline

Bagetta V, Picconi B, Marinucci S, Sgobio C, Pendolino V, Ghiglieri V, Fusco FR, Giampà C, Calabresi P (2011) Dopamine-dependent long-term depression is expressed in striatal spiny neurons of both direct and indirect pathways: implications for Parkinson's disease. J Neurosci 31:12513–12522. CrossRef Medline

Baker SN (2007) Oscillatory interactions between sensorimotor cortex and the periphery. Curr Opin Neurobiol 17:649–655. CrossRef Medline

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141. CrossRef Medline

Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Brain Res Rev 28:309–369. CrossRef Medline

Bourdy R, Barrot M (2012) A new control center for dopaminergic systems: pulling the VTA by the tail. Trends Neurosci 35:681–690. CrossRef Medline

Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68:815–834. CrossRef Medline

Brown P (2007) Abnormal oscillatory synchronisation in the motor system leads to impaired movement. Curr Opin Neurobiol 17:656–664. CrossRef Medline

Cabello N, Gandía J, Bertarelli DC, Watanabe M, Lluís C, Franco R, Ferré S, Luján R, Ciruela F (2009) Metabotropic glutamate type 5, dopamine D2 and adenosine A2a receptors form higher-order oligomers in living cells. J Neurochem 109:1497–1507. CrossRef Medline

Chan CS, Peterson JD, Gertler TS, Glajch KE, Quintana RE, Cui Q, Sebel LE, Plotkin JL, Shen W, Heiman M, Heintz N, Greengard P, Surmeier DJ (2012) Strain-specific regulation of striatal phenotype in Drd2-eGFP BAC transgenic mice. J Neurosci 32:9124–9132. CrossRef Medline

Chuhma N, Tanaka KF, Hen R, Rayport S (2011) Functional connectome of the striatal medium spiny neuron. J Neurosci 31:1183–1192. CrossRef Medline

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482:85–88. CrossRef Medline

Crittenden JR, Graybiel AM (2011) Basal ganglia disorders associated with imbalances in the striatal striosome and matrix compartments. Front Neuroanat 5:59. CrossRef Medline

Cunha RA (2001) Adenosine as a neuromodulator and as a homeostatic regulator in the nervous system: different roles, different sources and different receptors. Neurochem Int 38:107–125. CrossRef Medline

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298. CrossRef Medline

DeLong MR (1990) Primate models of movement disorders of basal ganglia origin. Trends Neurosci 13:281–285. CrossRef Medline

Enomoto K, Matsumoto N, Nakai S, Satoh T, Sato TK, Ueda Y, Inokawa H, Haruno M, Kimura M (2011) Dopamine neurons learn to encode the long-term value of multiple future rewards. Proc Natl Acad Sci U S A 108:15462–15467. CrossRef Medline

Flores-Hernández J, Cepeda C, Hernández-Echeagaray E, Calvert CR, Jokel ES, Fienberg AA, Greengard P, Levine MS (2002) Dopamine enhancement of NMDA currents in dissociated medium-sized striatal neurons: role of D1 receptors and DARPP-32. J Neurophysiol 88:3010–3020. CrossRef Medline

Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17:51–72. CrossRef Medline

Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306:1940–1943. CrossRef Medline

Fujisawa S, Buzsáki G (2011) A 4 Hz oscillation adaptively synchronizes prefrontal, VTA, and hippocampal activities. Neuron 72:153–165. CrossRef Medline

Fujiyama F, Sohn J, Nakano T, Furuta T, Nakamura KC, 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. CrossRef Medline

Gerfen CR, Surmeier DJ (2011) Modulation of striatal projection systems by dopamine. Annu Rev Neurosci 34:441–466. CrossRef Medline

Gerfen CR, Baimbridge KG, Miller JJ (1985) The neostriatal mosaic: compartmental distribution of calcium-binding protein and parvalbumin in the basal ganglia of the rat and monkey. Proc Natl Acad Sci U S A 82:8780–8784. CrossRef Medline

Gittis AH, Nelson AB, Thwin MT, Palop JJ, Kreitzer AC (2010) Distinct roles of GABAergic interneurons in the regulation of striatal output pathways. J Neurosci 30:2223–2234. CrossRef Medline

Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A 108 [Suppl 3]:15647–15654. CrossRef

Golding NL, Staff NP, Spruston N (2002) Dendritic spikes as a mechanism for cooperative long-term potentiation. Nature 418:326–331. CrossRef Medline

Graybiel AM (2005) The basal ganglia: learning new tricks and loving it. Curr Opin Neurobiol 15:638–644. CrossRef Medline

Hernández-Echeagaray E, Starling AJ, Cepeda C, Levine MS (2004) Modulation of AMPA currents by D2 dopamine receptors in striatal medium-sized spiny neurons: are dendrites necessary? Eur J Neurosci 19:2455–2463. CrossRef Medline

Higley MJ, Sabatini BL (2010) Competitive regulation of synaptic Ca^{2+} influx by D2 dopamine and A2A adenosine receptors. Nat Neurosci 13:958–966. CrossRef Medline

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

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

Hong S, Hikosaka O (2008) The globus pallidus sends reward-related signals to the lateral habenula. Neuron 60:720–729. CrossRef Medline

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. CrossRef Medline

Hong S, Jhou TC, Smith M, Saleem KS, Hikosaka O (2011) Negative reward signals from the lateral habenula to dopamine neurons are mediated by rostromedial tegmental nucleus in primates. J Neurosci 31:11457–11471. CrossRef Medline

Jhou TC, Geisler S, Marinelli M, Degarmo BA, Zahm DS (2009a) The mesopontine rostromedial tegmental nucleus: A structure targeted by the lateral habenula that projects to the ventral tegmental area of Tsai and substantia nigra compacta. J Comp Neurol 513:566–596. CrossRef Medline

Jhou TC, Fields HL, Baxter MG, Saper CB, Holland PC (2009b) The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. Neuron 61:786–800. CrossRef Medline

Kakade S, Dayan P (2002) Dopamine: generalization and bonuses. Neural Netw 15:549–559. CrossRef Medline

Kawagoe R, Takikawa Y, Hikosaka O (2004) Reward-predicting activity of dopamine and caudate neurons—a possible mechanism of motivational control of saccadic eye movement. J Neurophysiol 91:1013–1024. CrossRef Medline

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M (2007) Genetically determined differences in learning from errors. Science 318:1642–1645. CrossRef Medline

Kramer PF, Christensen CH, Hazelwood LA, Dobi A, Bock R, Sibley DR, Mateo Y, Alvarez VA (2011) Dopamine D_2 receptor overexpression alters behavior and physiology in Drd2–EGFP mice. J Neurosci 31:126–132. CrossRef Medline

Kravitz AV, Tye LD, Kreitzer AC (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci 15:816–818. CrossRef Medline

Kuramoto E, Furuta T, Nakamura KC, Unzai T, Hioki H, Kaneko T (2009) Two types of thalamocortical projections from the motor thalamic nuclei of the rat: a single neuron-tracing study using viral vectors. Cereb Cortex 19:2065–2077. CrossRef Medline

Lammel S, Lim BK, Ran C, Huang KW, Betley MJ, Tye KM, Deisseroth K, Malenka RC (2012) Input-specific control of reward and aversion in the ventral tegmental area. Nature 491:212–217. CrossRef Medline

Larkum ME, Zhu JJ, Sakmann B (1999) A new cellular mechanism for coupling inputs arriving at different cortical layers. Nature 398:338–341. CrossRef Medline

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463. CrossRef Medline

Legenstein R, Maass W (2011) Branch-specific plasticity enables self-organization of nonlinear computation in single neurons. J Neurosci 31:10787–10802. CrossRef Medline

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. CrossRef Medline

Levine MS, Li Z, Cepeda C, Cromwell HC, Altemus KL (1996) Neuromodulatory actions of dopamine on synaptically-evoked neostriatal responses in slices. Synapse 24:65–78. CrossRef Medline

Li X, Morita K, Robinson HP, Small M (2013) Control of layer 5 pyramidal cell spiking by oscillatory inhibition in the distal apical dendrites: a computational modelling study. J Neurophysiol. Advance online publication. Retrieved April 1, 2013. doi:10.1152/jn.00397.2012. CrossRef

Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 447:1111–1115. CrossRef Medline

Matsumoto M, Hikosaka O (2009) Representation of negative motivational value in the primate lateral habenula. Nat Neurosci 12:77–84. CrossRef Medline

McClure SM, Daw ND, Montague PR (2003) A computational substrate for incentive salience. Trends Neurosci 26:423–428. CrossRef Medline

Mena-Segovia J, Bolam JP, Magill PJ (2004) Pedunculopontine nucleus and basal ganglia: distant relatives or part of the same family? Trends Neurosci 27:585–588. CrossRef Medline

Meng Z, Li Q, Martin JH (2004) The transition from development to motor control function in the corticospinal system. J Neurosci 24:605–614. CrossRef Medline

Mitrano DA, Pare JF, Smith Y (2010) Ultrastructural relationships between cortical, thalamic, and amygdala glutamatergic inputs and group I metabotropic glutamate receptors in the rat accumbens. J Comp Neurol 518:1315–1329. CrossRef Medline

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947. Medline

Morishima M, Kawaguchi Y (2006) Recurrent connection patterns of corticostriatal pyramidal cells in frontal cortex. J Neurosci 26:4394–4405. CrossRef Medline

Morishima M, Morita K, Kubota Y, Kawaguchi Y (2011) Highly differentiated projection-specific cortical subnetworks. J Neurosci 31:10380–10391. CrossRef Medline

Morita K (2009) Computational implications of cooperative plasticity induction at nearby dendritic sites. Sci Signal 2:pe2. CrossRef Medline

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. CrossRef Medline

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057–1063. CrossRef Medline

Moyer JT, Wolf JA, Finkel LH (2007) Effects of dopaminergic modulation on the integrative properties of the ventral striatal medium spiny neuron. J Neurophysiol 98:3731–3748. CrossRef Medline

Murayama K, Matsumoto M, Izuma K, Matsumoto K (2010) Neural basis of the undermining effect of monetary reward on intrinsic motivation. Proc Natl Acad Sci U S A 107:20911–20916. CrossRef Medline

Nakamura K, Hikosaka O (2006) Role of dopamine in the primate caudate nucleus in reward modulation of saccades. J Neurosci 26:5360–5369. CrossRef Medline

Nakano T, Doi T, Yoshimoto J, Doya K (2010) A kinetic model of dopamine- and calcium-dependent striatal synaptic plasticity. PLoS Comput Biol 6:e1000670. CrossRef Medline

Nelson AB, Hang GB, Grueter BA, Pascoli V, Luscher C, Malenka RC, Kreitzer AC (2012) A comparison of striatal-dependent behaviors in wild-type and hemizygous Drd1a and Drd2 BAC transgenic mice. J Neurosci 32:9119–9123. CrossRef Medline

Niv Y (2007) Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? Ann N Y Acad Sci 1104:357–376. CrossRef Medline

Niv Y, Daw ND, Joel D, Dayan P (2007) Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) 191:507–520. CrossRef Medline

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454. CrossRef Medline

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. CrossRef Medline

Olveczky BP, Andalman AS, Fee MS (2005) Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. PLoS Biol 3:e153. CrossRef Medline

Paladini CA, Celada P, Tepper JM (1999) Striatal, pallidal, and pars reticulata evoked inhibition of nigrostriatal dopaminergic neurons is mediated by GABA(A) receptors in vivo. Neuroscience 89:799–812. CrossRef Medline

Parthasarathy HB, Graybiel AM (1997) Cortically driven immediate-early gene expression reflects modular influence of sensorimotor cortex on identified striatal neurons in the squirrel monkey. J Neurosci 17:2477–2491. Medline

Paton JJ, Louie K (2012) Reward and punishment illuminated. Nat Neurosci 15:807–809. CrossRef Medline

Pennartz CM, Ameerun RF, Groenewegen HJ, Lopes da Silva FH (1993) Synaptic plasticity in an in vitro slice preparation of the rat nucleus accumbens. Eur J Neurosci 5:107–117. CrossRef Medline

Pennartz CM, Ito R, Verschure PF, Battaglia FP, Robbins TW (2011) The hippocampal-striatal axis in learning, prediction and goal-directed behavior. Trends Neurosci 34:548–559. CrossRef Medline

Poirazi P, Mel BW (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron 29:779–796. CrossRef Medline

Polsky A, Mel B, Schiller J (2009) Encoding and decoding bursts by NMDA spikes in basal dendrites of layer 5 pyramidal neurons. J Neurosci 29:11891–11903. CrossRef Medline

Potjans W, Diesmann M, Morrison A (2011) An imperfect dopaminergic error signal can drive temporal-difference learning. PLoS Comput Biol 7:e1001133. CrossRef Medline

Reiner A, Hart NM, Lei W, Deng Y (2010) Corticostriatal projection neurons—dichotomous types and dichotomous functions. Front Neuroanat 4:142. CrossRef Medline

Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67–70. CrossRef Medline

Robbins TW, Everitt BJ (1996) Neurobehavioural mechanisms of reward and motivation. Curr Opin Neurobiol 6:228–236. CrossRef Medline

Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615–1624. CrossRef Medline

Rummery GA, Niranjan M (1994) On-line Q-learning using connectionist systems. In: Technical report CUED/F-INFENG/TR 166, Cambridge University Engineering Department. Cambridge, UK: Cambridge UP.

Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054–1069. CrossRef Medline

Salamone JD, Correa M (2002) Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behav Brain Res 137:3–25. CrossRef Medline

Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340. CrossRef Medline

Schiffmann SN, Fisone G, Moresco R, Cunha RA, Ferré S (2007) Adenosine A2A receptors and basal ganglia physiology. Prog Neurobiol 83:277–292. CrossRef Medline

Schiller J, Schiller Y (2001) NMDA receptor-mediated dendritic spikes and coincident signal amplification. Curr Opin Neurobiol 11:343–348. CrossRef Medline

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599. CrossRef Medline

Shabel SJ, Proulx CD, Trias A, Murphy RT, Malinow R (2012) Input to the lateral habenula from the basal ganglia Is excitatory, aversive, and suppressed by serotonin. Neuron 74:475–481. CrossRef Medline

Sheets PL, Suter BA, Kiritani T, Chan CS, Surmeier DJ, Shepherd GM (2011) Corticospinal-specific HCN expression in mouse motor cortex: I(h)-dependent synaptic integration as a candidate microcircuit mechanism involved in motor control. J Neurophysiol 106:2216–2231. CrossRef Medline

Shen W, Flajolet M, Greengard P, Surmeier DJ (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848–851. CrossRef Medline

Shen X, Ruan X, Zhao H (2012) Stimulation of midbrain dopaminergic structures modifies firing rates of rat lateral habenula neurons. PLoS One 7:e34323. CrossRef Medline

Silberberg G, Markram H (2007) Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells. Neuron 53:735–746. CrossRef Medline

Surmeier DJ, Carrillo-Reid L, Bargas J (2011) Dopaminergic modulation of striatal neurons, circuits, and assemblies. Neuroscience 198:3–18. CrossRef Medline

Sutton R, Barto A (1998) Reinforcement learning. Cambridge, MA: Massachusetts Institute of Technology.

Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'Donnell P, Niv Y, Schoenbaum G (2011) Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat Neurosci 14:1590–1597. CrossRef Medline

Takikawa Y, Kawagoe R, Hikosaka O (2004) A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. J Neurophysiol 92:2520–2529. CrossRef Medline

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887–893. CrossRef Medline

Tepper JM, Lee CR (2007) GABAergic control of substantia nigra dopaminergic neurons. Prog Brain Res 160:189–208. CrossRef Medline

Threlfell S, Lalic T, Platt NJ, Jennings KA, Deisseroth K, Cragg SJ (2012) Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron 75:58–64. CrossRef Medline

Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. Nature 398:704–708. CrossRef Medline

Ueta Y, Otsuka T, Morishima M, Ushimaru M, Kawaguchi Y (2013) Multiple layer 5 pyramidal cell subtypes relay cortical feedback from secondary to primary motor areas in rats. Cereb Cortex. Advance online publication. Retrieved April 3, 2013. doi:10.1093/cercor/bht088. CrossRef

van Seijen H, van Hasselt H, Whiteson S, Wiering M (2009) A theoretical and empirical analysis of expected sarsa. Proc IEEE Symp Adapt Dyn Program Reinforce Learn 2009:177–184. CrossRef

Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36:955–968. CrossRef Medline

Watabe-Uchida M, Zhu L, Ogawa SK, Vamanrao A, Uchida N (2012) Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron 74:858–873. CrossRef Medline

Watanabe K, Hikosaka O (2005) Immediate changes in anticipatory activity of caudate neurons associated with reversal of position-reward contingency. J Neurophysiol 94:1879–1887. CrossRef Medline

Watkins C (1989) Learning from delayed rewards. PhD thesis, University of Cambridge.