Distinct corticostriatal compartments drive competition between adaptive and automatized behavior

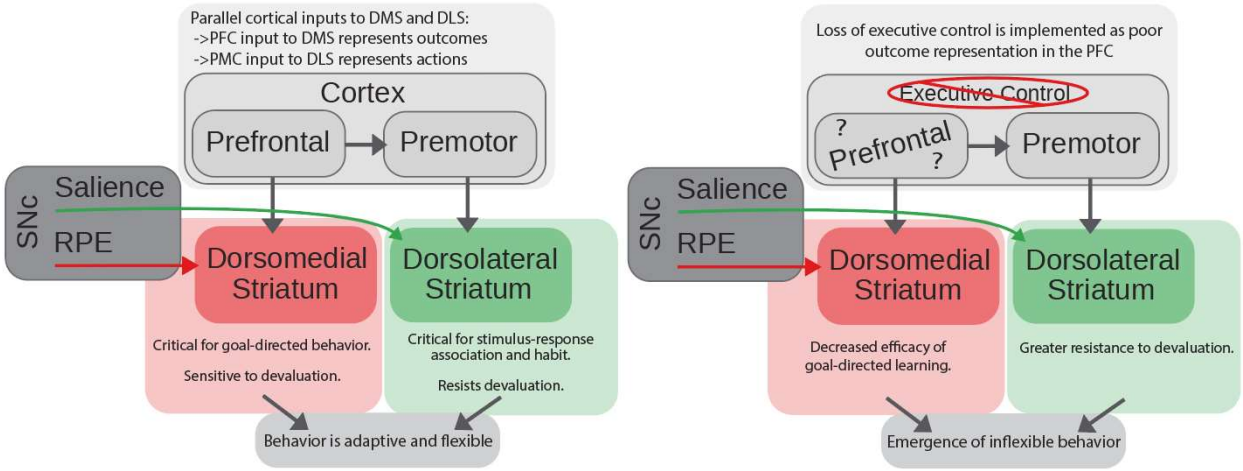
William H. Barnett[1], Alexey Kuznetsov[2], Christopher C. Lapish[1,3]

[1]Department of Psychology, Indiana University—Purdue University Indianapolis, Indianapolis, IN, 46202
[2]Department of Mathematics, Indiana University—Purdue University Indianapolis, Indianapolis, IN, 46202
[3]Stark Neurosciences Research Institute, Indiana University—Purdue University Indianapolis, Indianapolis, IN, 46202

Abstract

Cortical and basal ganglia circuits play a crucial role in the formation of goal-directed and habitual behaviors. In this study, we investigate the cortico-striatal circuitry involved in learning and how cortical interactions with specific striatal subregions are involved in the emergence of inflexible behaviors such those observed in addiction. Specifically, we develop a computational model of cortico-striatal interactions that performs concurrent goal-directed and stimulus-response learning. The model accomplishes this by distinguishing learning processes in the dorsomedial striatum (DMS) that rely on reward prediction error signals as distinct from the dorsolateral striatum (DLS) where learning is supported by salience signals. These striatal subregions each operate on unique cortical input: the DMS receives input from the prefrontal cortex (PFC), which represented outcomes and the DLS receives input from the premotor cortex which determined action selection. Following an initial learning of a two-alternative forced choice task, we subjected the model to reversal learning, reward devaluation, and punishment learning. Behavior driven by stimulus-response associations in the DLS resisted goal-directed learning of new reward feedback rules despite devaluation or punishment, indicating the expression of habit. We repeated these simulations after the loss of executive control, which was implemented as poor outcome representation in the PFC. Following this manipulation, no detectable effect of reward devaluation was observed, however, the efficacy of goal-directed learning was reduced, and stimulus response associations in the DLS were even more resistant to the learning of new reward feedback rules. In summary, this model describes how impaired coding in PFC reflects the loss of executive control thus leading to the emergence of inflexible behavior.

Graphical Abstract



Introduction

The repeated expression of a behavior is associated with the transition in the control of behavior from neural circuits that are optimized for flexible responding to those that are optimized for inflexible responding. It is thought that this transition involves the activation of neural circuitry in the basal ganglia that is devoted to habit and automaticity (Graybiel, 2008; Lipton et al., 2019). Molecular changes resulting from chronic substance abuse alter the role of this circuitry from enabling the performance of routine tasks without attention or cognitive load to driving compulsive drug seeking (Lüscher and Janak, 2021; Lüscher et al., 2020). Clearly articulating the computational processes that unfold across these circuits is critical for understanding the neural basis of behavioral control. Furthermore, it is critical to determine how these processes are altered in disorders that reflect the loss of behavioral control, such as addiction.

Here, we focus on two structures within the basal ganglia—specifically, the dorsomedial striatum (DMS) and dorsolateral striatum (DLS)—that have each been implicated in the pathophysiology of inflexible behavior and drug addiction (Corbit and Janak, 2016; Lipton et al., 2019). In an alcohol-seeking operant task, seeking behavior can be disrupted by inactivation of the DMS but is insensitive to inactivation of the DLS during early training (Corbit et al., 2012). After extensive training, the specificity of these manipulations becomes reversed: alcohol seeking is disrupted by inactivation of the DLS and is

insensitive to inactivation of the DMS. Similarly, cocaine-seeking behavior is sensitive to disruption of the DMS during early learning and disruption of the DLS after extensive training (Murray et al., 2012). In summary, these partitions of the dorsal striatum have distinct temporal contributions to drug-seeking behavior where DMS is initially critical and eventually transitions to DLS.

The distinct roles of the DMS and DLS in inflexible behavior are derived from differences in their computational properties. The DMS and DLS are respectively involved in goal-directed and habitual behavior (Schwabe and Wolf, 2011). Manipulations that destroy or disrupt the DMS and DLS during instrumental behavioral tasks reveal their respective roles in goal-directed learning and the formation of habit (Yin and Knowlton, 2006). The inactivation of the DMS decreases sensitivity to reward devaluation (Yin et al., 2005), and the destruction of the DLS increases sensitivity to reward devaluation and abolishes habitual seeking in the absence of a reward (Yin et al., 2004). There is evidence that it takes more time to engage learning mechanisms in the DLS; plasticity in the DLS occurs relatively slowly compared to DMS as training progresses (Yin et al., 2009).

Differences in the computational function of the DMS and DLS are supported by differences in their synaptic plasticity mechanisms and differences in their cortical inputs. In the basal ganglia hypothesis for reward-based learning, action selection is gated in the striatum by dopamine-mediated synaptic plasticity of cortical projections to medium spiny neurons (Frank, 2005; Graybiel, 2008). In classical reinforcement learning, dopamine release encodes reward prediction error (RPE). However, there is evidence that different compartments of the striatum receive partition-specific nigro-striatal projections that encode different information (Lerner et al., 2015; Matsumoto and Hikosaka, 2009). Classical RPE-encoding nigrostriatal neurons project to the DMS, and salience-encoding neurons project to the DLS (Lerner et al., 2015). Moreover, dopamine release in the basal ganglia acts on cortico-striatal synapses, and cortical input to the striatum is topographically organized across the DMS and DLS. Inputs to the DS exhibit a clear topographic bias where the medial portion of the striatum is more likely to

receive input from the prefrontal cortex (PFC), and the lateral portion more likely to receive input from somatosensory and motor regions (Hunnicutt et al., 2016; Pan et al., 2010; Peters et al., 2021).

The input from PFC to the DMS is of particular interest. In rodents, the PFC is important for executive function and goal-directed learning (Barker et al., 2015; Hart et al., 2018a, 2018b; Kesner and Churchwell, 2011; Ostlund and Balleine, 2005; Tran-Tu-Yen et al., 2009). Anatomically comparable structures in humans and non-human primates are involved in cognitive and executive function as well as goal-directed behavior (Balleine and O'Doherty, 2010; Donahue and Lee, 2015; Kennerley et al., 2009; Laubach et al., 2018; Perry et al., 2011; Tsutsui et al., 2016). Critically, these regions are impaired following prolonged alcohol use (Schacht et al., 2013) and this impairment corresponds to increased responding to alcohol (Crews and Boettiger, 2009; Myrick et al., 2004). In recent work, we investigated how the computational properties of the PFC are altered in a rodent model of excessive alcohol consumption and neuronal activity in the medial PFC fails to appropriately code for intent to drink and seeking behavior (Linsenbardt et al., 2019; Timme et al., 2021). Taken in combination with observations about the DMS and DLS, we leverage these results to illustrate a hypothesis for how impairment in the PFC contributes to the emergence of inflexible behavior.

Here we present a new computational model of cortico-striatal learning that incorporates goal-directed learning in the DMS and stimulus-response learning in the DLS. We derive this implementation from a theory of reinforcement learning based on dopamine-mediated plasticity of cortico-striatal projections to medium spiny neurons (MSNs) (Graybiel, 2008). Dopamine induces long-term potentiation and long-term depression in D1 and D2 receptor expressing MSNs respectively, and these changes are hypothesized to configure the basal ganglia to selectively disinhibit thalamocortical relay neurons in the context of ongoing behavioral tasks (Frank, 2005). This hypothesis has been incorporated in computational models that include reward-based learning that is based on RPE in the striatum (Frank, 2005; Kim et al., 2017; Mulcahy et al., 2020). However, stimulus-response or habit was previously

implemented as a consequence of Hebbian learning in cortico-cortical projections (Kim et al., 2017; Mulcahy et al., 2020). Recent publications have investigated contemporaneous goal-directed and habitual learning in a model of reinforcement learning (Miller et al., 2019; Tsutsui-Kimura et al., 2020) and the consequences of the spatial distribution of dopamine release in the medial-lateral axis of the dorsal stratum (Hamid et al., 2021). To understand the mechanism of how behaviors transition from goal-directed to habitual, a computational model is required that is capable of articulating how changes in cortico-striatal plasticity support both goal-directed learning in the DMS and stimulus-response learning in the DLS, which we provide here for the first time.

In the present study, we simulate two-alternative forced choice behavioral tasks to investigate the interaction of the DMS and the DLS. Following an initial learning session, we challenge the model with reward devaluation, reward reversal, and punishment learning. These different behavioral tasks are implemented by manipulations to the magnitude, action contingency, and valence of the reward feedback. We challenge the model again in scenarios characterized by the loss of executive control. In these simulations, neural activity of the PFC fails to appropriately code for action-selection. Model performance is quantified by calculating the likelihood of action selection on a trial-by-trial basis over an ensemble of model instances. Our results demonstrate how the loss of executive function could reduce the efficacy of goal-directed learning and emphasize the expression of stimulus-response or habitual behavior.

Results

Organization of cortico-striatal partitions.

In this study, we extend our previous model of the basal ganglia to implement a neural network that combines learning in the DMS and learning in the DLS to perform a two-alternative forced-choice decision-making task (Fig. 1A). This new model implements the pattern of choice-specific channels for

neuronal circuitry (Frank, 2005; Kim et al., 2017; Mulcahy et al., 2020) to perform reward-based learning across partitions of the dorsal striatum (Lerner et al., 2015; Matsumoto and Hikosaka, 2009). The medial partition, which incorporates the PFC and the DMS, and the lateral partition, which incorporates the PMC and the DLS. Each partition also contains a direct and indirect pathway for each of the two possible actions. Each partition engages the machinery of the BG to utilize input from cortex, and the processing of this input results in inhibition back onto the cortex in a way that reflects the accumulation of past experience in cortico-striatal weights. By implementing different synaptic weight update rules in the DMS and DLS, different modes of learning are obtained over successive simulated forced-choice trials. The activity of the PFC (in the medial partition) at the end of a trial indicates the choice that results from cortico-striatal weights in the DMS that have been determined by the reward prediction error (RPE) (Fig. 1B). We conceptualize this aspect of the decision-making process as the formation of the representation a selected outcome. In the lateral partition, the PMC receives inhibitory input from the BG that results from cortico-striatal weights in the DLS. In contrast to the DMS, cortico-striatal weights in the DLS have been informed by expected reward. The PMC also receives excitatory input from the PFC. In this way, the PMC integrates inputs that are associated with both reward prediction error and expected reward (Fig. 1B). The activity of the PMC at the end of a trial defines the behavioral output of the system. We conceptualize this aspect of the decision-making process as the formation of an action representation.

Organization of the basal ganglia.

In our model, cortical input to the striatum represents environmental state and the deliberation of choice selection. The BG is organized into two subsequent pathways that descend from the striatum: the direct pathway (striatum-SNr/GPi) and the indirect pathway (striatum-GPe-STN-SNr/GPi) (Fig. 1A). The direct and indirect pathways converge on the SNr and the GPi, which are the output node of the BG;

here, these two regions are treated as the same populations. D1 MSNs directly inhibit the SNr/GPi. D2 MSNs inhibit the GPe which subsequently inhibits the SNr/GPi; in other words, D2 MSN activation leads to disinhibition of the SNr/GPi. Preferential activation of the direct pathway promotes the selection of an action. When D1 MSNs are active, they suppress SNr/GPi neurons that would otherwise inhibit the representation of that action in the cortex. On the other hand, preferential activation of the indirect pathway biases against the selection of that action. D2 MSN activation releases SNr/GPi from inhibition, and the corresponding cortical targets are suppressed. Learning in the BG is contingent on the release of dopamine (DA) from the SNc; this reward-based signal induces long-term potentiation in D1 MSNs and long-term depression in D2 MSNs. By altering the ability of D1 and D2 MSNs to utilize cortical input, DA release supports reward-based learning in this system.

A biophysical model of the basal ganglia that includes both DMS and DLS.

Deliberation and selection between choices occurred concurrently within the medial and lateral partitions, and we distinguished the results of these two separate processes based on learning modality. Since learning in the medial partition was based on reward prediction error, the PFC-DMS weights were updated after each trial to accumulate information about reward contingency (Fig. 1B). As such, we characterized learning in the medial partition as goal-directed learning, and we designated the choice made within the medial partition to be outcome selection. On the other hand, learning in the lateral partition was based on expected reward. This signal contained information about recent reward history, and the update of PMC-DLS synaptic weights accumulated information about the likelihood that a given stimulus would result in a reward (Fig. 1B). Learning in the lateral partition associated stimuli to rewarding responses. Since the output of the lateral partition defined behavior, we designated the choice made in the lateral partition to be action selection.

The deliberation that determined outcome selection was independent of the deliberation that determined action selection, and depending on context and history, these processes may agree or disagree (Fig. 1C,D). The firing rate of the PFC population determined outcome selection, and each PFC population excited the corresponding PMC population. In this way, outcome selection influenced action selection in the lateral partition. If the PFC population corresponding to outcome #1 was selected, the high firing rate of that PFC population made it more likely that action #1 was selected in the lateral partition, and in some cases both partitions agreed: outcome #1 and action #1 were both selected (Fig. 1C). However, deliberation in the lateral partition was also subject to its own learned stimulus-response associations, and on occasion, the internal dynamics of the lateral partition competed with input from the medial partition to determine which action is performed. As a result, there were cases where the two partitions disagreed: PFC population #2 had a high firing rate indicating the selection of outcome #2, but in the lateral partition, PMC #1 had a high firing rate indicating the selection of action #1 (Fig. 1D).

Reward engages goal-directed learning and ultimately the formation of a stimulus-response association.

Learning in this model was accomplished by changes in the cortico-striatal synaptic weights. Generally, the enhancement of D1 synaptic weights promoted the selection of the corresponding action, and the enhancement of D2 synaptic weights made the selection of the corresponding action less likely. At the beginning of the initial learning session, D1 and D2 weights were not biased to select a particular action (Fig. 2A; Fig. 2B1 marker $*1$). Over many trials, the reward feedback from the selected action was used to update these weights such that one action was more or less likely to be selected. As discussed below, the medial and lateral partitions utilized this reward feedback in different ways.

In the medial partition, the RPE was used to update weights. This quantity could be either positive or negative and change sign from one trial to the next. For example, when there was a low expected reward and the rewarded action (in this case action #1) was chosen, the RPE became positive

(Fig. 2A; Fig. 2B2 marker ∗2). If the PFC population representing outcome #1 was active, the PFC-DMS D1 weight for outcome #1 was enhanced and the PFC-DMS D2 weight for outcome #1 was suppressed (Fig. 2A; Fig. 2B3 marker ∗3). Since PFC population #2 was not active at this step, PFC-DMS weights for outcome #2 were not updated based on RPE; these weights slowly decayed to their baseline value.

On the other hand, if the unrewarded action (action #2) was chosen, the RPE became negative (Fig. 2A; Fig. 2B4 marker ∗4). If the outcome #1 neurons were also activated in this case, the PFC-DMS D1 weight for action #1 was suppressed and the PFC-DMS D2 weight for action #1 was enhanced (Fig. 2A; Fig. 2B4-5 marker ∗5). However, if the unrewarded action (action #2) is selected and the PFC selected outcome #2, the PFC-DMS weights were changed to de-emphasize the selection of outcome #2: the D1 weight was decreased, and the D2 weight was increased (Fig. 2A; Fig. 2B4-5 marker ∗6). Over the course of training, the PFC-DMS weights were updated such that the DMS channel for outcome #1 became good at utilizing cortical input to activate the direct pathway, which emphasized the selection of outcome #1, and the DMS channel for action #2 became good at utilizing cortical input to activate the indirect pathway, which tended to prevent the selection of outcome #2. Thus, the cortico-striatal weights in the DMS embodied reward contingency.

In the lateral partition, the expected reward was used to update weights. In this session, expected reward was strictly positive. When the rewarded action (action #1) was chosen, the PMC-DLS D1 weight for that action was enhanced and the PMC-DLS D2 weight for that action was suppressed (Fig. 2A marker ∗7). When the unrewarded action was chosen, the expected reward decreased but remained positive. The PMC-DLS D1 weight was also enhanced and its PMC-DLS D2 weight was also suppressed (Fig. 2A marker ∗8). The differentiation of PMC-DLS D1 and PMC-DLS D2 weights could become biased across channels only when one action is selected many more times than the other action. Since the agent selected action #1 more frequently than action #2, the PMC-DLS D1 and PMC-DLS D2 weights for action #1 ultimately became biased to select for action #1. Since action #2 was less frequently selected,

the weight update between trials was dominated by the decay term, and these weights stayed near their baseline value. In this way, the DLS learns a stimulus-response association for the rewarded action but has accumulated no information about the unrewarded action. Behavior driven by this stimulus-response association is determined by the frequency at which this action was rewarded rather than contingency upon outcome.

Reward reversal engages goal directed learning to promote a new stimulus-response association.

The reversal learning session began with cortico-striatal synaptic weights inherited from the end of the initial learning session; this model instance was already biased to select action #1 (Fig. 3A). In this session, the reward feedback was altered such that action #2 was the rewarded action. In early trials, the PMC continued to select the previously rewarded action #1 due to the persistence of PMC-DLS synaptic weights that favored action #1. The expected reward was high, and when action #1 was selected, the reward prediction error became negative. This selection suppressed the PFC-DMS D1 weight for action #1, enhanced the PFC-DMS D2 weight for action #1, and led to a decrease in the expected reward (Fig. 3A; Fig. 3B1 marker $*1$). However, the PFC-DMS synaptic weights for action #2 required the selection of action #2 to be actively updated. Since the PMC-DLS weights were biased to select action #1, there was a low likelihood of selecting action #2 for a number of trials. During these early trials, synaptic weights in the DMS became biased to favor the indirect pathway for both action #1 and action #2 (Fig. 3A; Fig. 3B2 marker $*2$), and there was not a strong likelihood for the DMS to emphasize one action over the other. The medial partition does not receive input from the lateral partition, so it is free to produce an outcome selection based on its own internal dynamics. Since synaptic weights in the DMS were biased to promote both indirect pathways, the PFC was likely to select either outcome #1 or outcome #2 for several trials. In other words, the PFC performed exploration of outcomes. Once the rewarded action #2 was selected, the corresponding PFC-DMS D1 synaptic weight was enhanced; the corresponding PFC-DMS D2 synaptic weight was suppressed; and the subsequent

selection rate for outcome #2 increased in the medial partition (Fig. 3A; Fig. 3B3 marker ∗3). The combination of occasional excitatory input to the PMC resulting from the selection of outcome #2 in the PFC and the gradual decay of PMC-DLS synaptic weights that favored action #1 lead to a transition in PMC action selection from the persistent selection of action #1 to exploration of action selection.

The decay of PMC-DLS synaptic weights towards the baseline is due to a decrease of the expected reward in early trials. The channel for action #1 in the DLS was less likely to promote action #1, and the decay of these weights contributed to exploration of actions in the lateral partition (Fig. 3A marker ∗4). Concomitantly, goal-directed selection of outcome #2 in the PFC promoted the selection of action #2 in the PMC. During this exploratory phase, the occasional selection of action #2 increased the expected reward, and the PMC-DLS weights for action #2 began to differentiate—the PMC-DLS D1 weight for action #2 increased and the PMC-DLS D2 weight for action #2 decreased—such that the selection of action #2 was promoted (Fig. 3A marker ∗5). In this way, the DLS lost the stimulus-response association for action #1 and learned a stimulus-response association for action #2.

Reward devaluation decreases selection accuracy.

The devaluation session began with cortico-striatal weights inherited from the end of the initial learning session; this model instance was already biased to select action #1. In this session, we decreased the reward feedback for action #1 to 20% of the reward feedback value in the initial learning session. This is based on the observation that the dopamine response scales with the magnitude of the reward (Lak et al., 2014; Tobler et al., 2005).

In this session, the expected reward began close to 1, however, the selection of action #1 provided decreased reward feedback. The expected reward was large compared to the reward feedback, and the RPE became negative in early trials despite the selection of the rewarded action (Fig. 4A; Fig. 4B1 marker ∗1). Since the RPE was negative, PFC-DMS D1 weights for action #1 were

suppressed and PFC-DMS D2 weights for action #1 were enhanced in early trials: the channel for action #1 in the medial partition was biased to select against outcome #1 (Fig. 4A; Fig. 4B2 marker ∗2). However, the selection of action #2 was not rewarded; there was no change in the valence of PFC-DMS weights for action #2. The agent was less likely to select outcome #1, but it was still more likely to select outcome #1 over outcome #2. As it integrated the newly reduced reward feedback, the expected reward converged to the new basin of attraction; in these trials, action selection pushed the expected reward close to 0.2, and mistakes in action selection reduced the expected reward below 0.2. Once the expected reward followed this new regime, the relative exploratory phase in outcome selection was abolished. The RPE became positive again when the rewarded action was selected, and the PFC-DMS weights for action #1 were restored to select for action #1 (Fig. 4A; Fig. 4B3 marker ∗3).

The change in magnitude of the reward altered the steady state value of PMC-DLS weights. Since the reward was diminished, the expected reward was also relatively small. The expected reward modulates the magnitude of change in PMC-DLS synaptic weights, so the amount by which action selection changed PMC-DLS weights was relatively small. This active update of synaptic weights was required to prevent decay of weights to their baseline values, and over many trials, the PMC-DLS D1 and D2 weights converged to new values that were closer to baseline synaptic weights. While the stimulus-response association for action #1 was diminished, it was not lost entirely. Despite reward devaluation, the agent still persistently selected action #1 with greater likelihood.

Punishment engages goal-directed learning.

The punishment session began with cortico-striatal weights inherited from the end of the initial learning session; this model instance was already biased to select action #1. In this session, the magnitude of the reward feedback for action #1 was 50% that in the initial learning session, but it had

the opposite valence. We define this negative reward to be punishment, which is based on experiments that show pauses in DA firing in response to aversive stimuli (Lerner et al., 2015).

In early trials of the punishment session, the PMC continued to select the previously rewarded action #1 due to the persistence of PMC-DLS synaptic weights that favored action #1 (Fig. 5A). In the first trial, the expected reward was close to 1. Selection of either action #1 or action #2 decreased the expected reward since the reward feedback was never positive. Similar to the early trials of the reward devaluation session, the RPE was strongly and consistently negative at the beginning of this session. For several trials, the expected reward was influenced by its initial vale, which was based on the initial training session. During these trials, the PFC-DMS D1 and D2 weights for outcome #1 changed valence such that the DMS channel for action #1 emphasized the indirect pathway over the direct pathway, and the PFC performed outcome exploration (Fig. 5A; Fig. 5B1 marker $*1$). During exploration, the PFC-DMS weights for both channels emphasized the indirect pathway (Fig. 5A; Fig. 5B2 marker $*2$). Finally, PFC-DMS weights for outcome #2 began to emphasize the direct pathway and suppress the indirect pathway as the PFC began to persistently select outcome #2 (Fig. 5A; Fig. 5B3 marker $*3$). The progression of PFC-DMS weights in this session was similar to the progression of weights in the reward reversal session.

In these simulations the lateral partition utilized the *rectified* reward expectation to perform weight updates. The rectified expectation integrates the absolute value of the reward feedback, which reflects salience of the selection option. In all previous sessions, the reward feedback was positive and the rectified expectation was identical to expected reward. However, in the punishment session, the expected reward becomes negative, and the rectified expectation diverged from the expected reward. The combination of decreased selection of action #1 and the reduced rectified expectation prevented the agent from maintaining PMC-DLS D1 and D2 weights for action #1 in such a way to emphasize this action. As these weights converged to their baseline values, the PMC selected the unpunished action #2 more preferentially than punished action #1 based on input from the PFC. However, the PMC-DLS

weights for action #2 did not strongly develop to select for action #2. Here, action selection was driven by the PFC which sought to minimize the delivery of punishment. This operation kept the rectified expectation close to zero. Since the update rule for PMC-DLS depends on rectified expectation, the change in PMC-DLS synaptic weights was not of great enough magnitude to overcome the decay of synaptic weights after each trial. Therefore, punishment learning does not develop the stimulus-response association of choosing an alternative action.

Misalignment of PFC outcome selection with channels in DMS and PMC decreases action selection performance.

Work from our group shows that outcome representations are reduced in animals that drink alcohol compulsively, and also demonstrate impaired behavioral flexibility (De Falco et al., 2021; Linsenbardt et al., 2019; Timme et al., 2021). More specifically, reductions in mutual information about two competing options was observed in neural recordings obtained from the mPFC of rats that drink alcohol compulsively, indicating that the representation of these options was not clearly disambiguated. To simulate this observation, we repeated reward reversal and punishment simulations incorporating reduced outcome coding in the PFC (Fig. 6). In these simulations, each PFC population routed 90% of its projections to the appropriate channel and 10% of its projections to the wrong channel (striatal neurons corresponding to the other outcome) (Fig. 6A). The inhibitory input from the GPi populations to the PFC was similarly mis-matched. This manipulation compromised the fidelity of the PFC signal as PFC activity did not correlate as well in these simulations to action selection. In simulations of both reward reversal and punishment, reduced outcome coding in the PFC was sufficient to handicap performance in action selection tasks (Fig. 6).

To characterize the change in agent performance due to PFC misalignment, we performed a change point analysis on outcomes and actions. In each model instance, change points in cortical activity

were detected as the first zero-crossing of the log likelihood ratio in a normative model of evidence accumulation. This analysis applied separately to PFC and PMC activity (Fig 7A) and yielded a trial number for the change point of both cortical compartments. For each behavioral session, we performed this analysis for 100 distinct agents (see Methods). We compared outcome selection and action selection change points between Reward Reversal and Reward Reversal with the misaligned PFC ($\mathrm{PFC}^{\dagger}$) sessions and between Punishment and Punishment $\mathrm{PFC}^{\dagger}$ sessions (Fig. 7B). From the Reward Reversal to the Reward Reversal $\mathrm{PFC}^{\dagger}$ sessions, the median agent change point increased from trial 15 to 20 in outcome in the PFC (Mann-Whitney U = 2933.5, n_1=n_2=100, p=4.3E-7) and from trial 47 to 82 in the PMC (Mann-Whitney U = 342.5, n_1=n_2=100, p=5.2E-30). From Punishment to the Punishment $\mathrm{PFC}^{\dagger}$ sessions, the median agent change point increased from trial 16.5 to 22 in the PFC (Mann-Whitney U = 3526.5, n_1=n_2=100, p=3.16E-4) and from trial 77 to 152.5 in the PMC (Mann-Whitney U = 34, n_1=n_2=100, p=7.0E-34). Thus, the simulations bridge the compromised outcome coding in the PFC and reduced performance in the action selection tasks.

Discrepancy between steady state performance of goal-directed and stimulus-response learning.

In several of the behavioral challenges, steady state agent performance differed between the PFC in the medial partition and the PMC in the lateral partition. Steady state performance was defined as the mean agent performance at the end of each behavioral session (Fig. 7C). In the initial learning session and in reward reversal sessions, steady state performance was greater in the PMC than the PFC; the PMC was more likely to choose the rewarded action than the PFC was to choose the corresponding outcome (Table 1). In the devaluation session, there was no significant difference between PFC and PMC performance (Table 1). In punishment sessions, the PMC was less likely to choose the unpunished action than the PFC was to choose the corresponding outcome (Table 1). The better performance in the PMC

reflects the accumulated contribution of the past choices encoded in the PMC-DLS weights, which indicates a learned stimulus-response association.

Session Name	Mann-Whitney U statistic	p-value
Initial Learning	1509.0	5.8E-17
Devaluation	4733.0	0.51
Reward Reversal	8425.5	5.8E-17
Punishment	2248.0	1.8E-11
Reward Reversal PFC†	7892.0	1.6E-12
Punishment PFC†	563.0	2.2E-27

Table 1. Discrepancy between goal-directed learning and habit. In each test, n_1=n_2=100. Comparisons can be visualized in the notched box-and-whisker plots on Fig. 7B.

Discussion

The main aim of this study was to formalize the computational mechanisms that support the expression of goal-directed and habitual behavior, as well as the transitions between them. Here, we showed how a stimulus-response association—embodied in the accumulation of cortico-striatal synaptic weights in the DLS—could prevent agents from quickly learning new reward feedback rules during simulated behavioral tasks. Indeed, DLS-driven action selection perseverated in the selection of the previously rewarded action even after cortico-striatal weights in the DMS were configured in a manner to support exploration. Specifically, DLS-driven action selection resisted goal-directed behavior despite reward devaluation and punishment, and in these scenarios, we characterized behavior as habitual.

We sought to implement a simple circuit-level manipulation in our computational model to illustrate the emergence of inflexible behavior through primacy of the DLS in action selection. The

expression of choices that were previously rewarded in the face of changing contingency is a hallmark of inflexible behavior, and the DLS is critical for the expression of inflexible behavior (Corbit et al., 2012; Lipton et al., 2019). On the other hand, the DMS is critical for goal-directed behavior, and disruptions to the dorsal striatum that favor the DMS may restore goal-directed behavior even in the face of compulsive seeking. In compulsive drinking, PFC neural firing rates code poorly for outcome selection (Linsenbardt et al., 2019; Timme et al., 2021). We developed a prediction for inflexible behavior based solely on PFC impairment. If the representation of outcomes in the PFC fail to correlate well with actions, then goal-directed learning in the DMS will lose efficacy, and agents will take longer to overcome previously learned stimulus-response associations. This loss of executive control was implemented in reward reversal and punishment learning tasks; agents with impaired PFC took longer to learn new reward feedback rules and exhibited worse steady state performance.

A recent investigation challenged model-free learning as a mechanism for habit in reinforcement learning (Miller et al., 2019). Instead, the authors propose a value-free habit mechanism to describe stimulus response association based on Thorndike's second Law of Exercise. Their mechanism for value-free learning is similar to the action of the DLS presented in the present report, though we have implemented our model as neural circuitry. Notably, the learning mechanisms described in (Miller et al., 2019) emphasize goal-directed behavior under variable-ratio reinforcement schedules and stimulus response behavior under variable-interval reinforcement schedules. This will be important to assess in future work with the model developed herein. To accomplish this it will require a mechanism to link the precise timing of the reward to the action performed, which is currently not implemented.

Reward devaluation and punishment

We distinguish between reward devaluation tasks—in which the magnitude of the reward is decreased, and punishment learning tasks—in which the valence of the reward feedback becomes negative. Experimental protocols may reduce value through manipulations to remove drive (i.e. selective satiety) or by inducing an aversion association (Rossi and Yin, 2012). Here, our model design was motivated in part by experiments that included a foot shock stimulus (Lerner et al., 2015). The difference between the implementation of these two behavioral tasks lies in the formulation of the reward feedback in the lateral partition. The reward feedback to the DLS—the rectified expectation—integrates the absolute value of the trial reward and deviates from the expected reward. In this way, the dopaminergic signal to the DLS is positive for either a rewarding or an aversive stimulus (as in Lerner et al., 2015). In goal-directed behavior, the aversive choice is avoided; in this case the agent seeks to minimize the negative reward prediction error and, hence, the quantity of negative expected reward. Since the rectified expectation is kept relatively small by the learning process, no new stimulus-response association is able to develop in punishment learning tasks. The results of punishment learning differ from the results of reward reversal. In both cases, the stimulus-response association for a previously rewarded action must be overcome. In the reward reversal session, goal-directed behavior works to maximize the reward delivered for the newly rewarded action. Hence, the expected reward is maintained at a large value over the course of training, and a new stimulus-response association is formed. This critical difference between reward reversal and punishment learning—whether there is a new stimulus-response association or not—is visualized in the comparison of steady state performance (Fig. 7B). In reward reversal simulations, steady state behavior is driven by the new stimulus-response association formed in the DLS, and the performance of outcome selection in the PFC is not as great as action selection in the PMC (discussed below). However, in punishment learning, steady state behavior is maintained by goal-directed behavior (the medial cortico-striatal partition) in the absence of a new stimulus-response association in the DLS. The performance of outcome selection in the PFC leads

performance of action selection in the PMC, which is slightly degraded by noise added during the simulation.

PFC coding failure

Our implementation of PFC impairment is intended to represent a failure of neural activity in the PFC to appropriately code for action selection. The cortical input to channels for outcome #1 and outcome #2 in the DMS and the PMC populations for action #1 and action #2 were slightly mixed in $\mathrm{PFC}^{\dagger}$ sessions. For example, the input to the DMS channel for outcome #1 and the input to the PMC population for action #1 was a weighted some of both PFC populations: 90% PFC for outcome #1 + 10% PFC for outcome #2 (Fig. 6A). During selection of an outcome in the PFC, the PFC population corresponding to the selected outcome had a high firing rate, and the PCF population corresponding to the unselected outcome had a low firing rate (Fig. 1 C,D). The firing rates of these populations at the end of the simulation were used to update cortico-striatal synaptic weights in the DMS between trials of the session (see Methods). Due to the manipulation, the firing rate used to update synaptic weights for the selected outcome is slightly lesser compared to control simulations, resulting in a slightly lesser change in synaptic weights. On the other hand, the firing rate used to update synaptic weights for the unselected outcome is slightly greater compared to control simulations, resulting in an inappropriately large change in synaptic weights. Similarly, the excitatory input from PFC to PMC was less distinct between channels: the PFC was slightly less able to promote the selection of the appropriate action in the PMC in $\mathrm{PFC}^{\dagger}$sessions compared to control simulations.

Our simulations of PFC impairment are intended to illustrate how a loss of executive control in the PFC could promote the emergence of inflexible behavior. How could repeated exposure to alcohol change the brain to induce a failure for neural activity to properly code for action selection? Dopamine dynamics in the PFC are critical for higher functions and executive control (Ott and Nieder, 2019; Puig et

al., 2014), and changes to dopamine modulation in the PFC are concurrent with alcohol-induced deficits in executive function (Trantham-Davidson and Chandler, 2015). Patients that suffer from a history of alcohol abuse may have deficient expression of D2 dopamine receptors in the PFC (Volkow et al., 1996, 2002). In rats, chronic exposure to alcohol disrupts D2 receptor function in the PFC (Trantham-Davidson et al., 2014). This may promote abnormal connectivity in the PFC, which, in particular, connects the neurons responsible for different outcomes and ultimately mixes the populations. It is thought that D1 receptors in the PFC act to stabilize network activity and D2 receptors promote flexibility (Durstewitz et al., 2000; Jenni et al., 2017). Disruptions in the PFC dopamine system that favor D1 receptors may render an individual unable to quickly adapt in the face of changing behavioral context.

The role and organization of direct and indirect pathways

In reversal learning and punishment follow-up sessions, we observe two prominent phases in the evolution of cortico-striatal weights in the DMS. First, the BG channel for the previously rewarded action becomes essentially inactivated. The cortico-striatal weights for D1 MSNs decreased in value below baseline, and the weights for D2 MSNs increased above baseline. During this phase, the model performance changed to become less likely to select the previously rewarded action, and as a result, it explored alternative actions—here there was only one alternative action. Second, this exploration phase was terminated by the potentiation of the cortico-striatal weight for D1 and depression of D2 MSNs that correspond to the newly rewarded action. This model differs from our previous BG computational models in that cortico-striatal weights do not rest at a baseline of zero (Kim et al., 2017; Mulcahy et al., 2020). In these zeroed resting weight models (ZRWMs), channel-specific feedback is required to engage the direct or indirect pathway; otherwise, a BG channel is not actively engaged in promoting or inhibiting itself during action-selection. In reversal tasks, the cortico-striatal weight for D2 MSNs corresponding to the previously rewarded action is increased from zero to engage that indirect pathway

in preventing the selection of the previously rewarded action, and the ZRWM proceeds in the exploration of alternative actions. In the present model, the cortico-striatal weights possess a baseline value, so for each channel, the direct and indirect pathways nominally promote or inhibit their own selection. The differentiation of D1 and D2 MSN weights alters the computations that a channel performs to favor promotion or inhibition of that channel. Despite differences in baseline cortico-striatal synaptic weights, the present model initiates exploration in a similar manner to ZRWMs: both mechanisms are based on the recruitment of the indirect pathway to inhibit the selection of the previously rewarded actions.

The observation that both direct and indirect pathway MSNs are activated at the onset of movement supports the hypothesis that direct and indirect pathways are concurrently involved in selection decisions (Cui et al., 2013; Tecuapetla et al., 2016). In the "complementary" model of the striatum, clusters of direct pathway MSNs are selectively activated while indirect pathway MSNs broadly inhibit competing clusters (Bariselli et al., 2019). In the present model, each action is represented by a cluster of direct pathway MSNs within the striatum. However, the present model differs from the "complementary" model in that each cluster of direct pathway MSNs acts in opposition to a selective cluster of indirect pathway MSNs. The computation performed in the SNr/GPi output node of our model is based on competition between the direct and indirect pathways associated with each action; in other words, the BG channel for each action emits a GO or a NOGO signal. In this way, our computational model is most like the "competitive" model described in (Bariselli et al., 2019) where direct and indirect pathway MSNs compete within an ensemble to promote or inhibit an action. In this study, our model realizes multiple such channels to represent distinct action options, and competition between channels—and hence between actions—is implemented via mutual inhibition within each cortical compartment.

Striatal Dopamine Distribution

The spatio-temporal distribution of dopamine release within the striatum is critical for selective reward credit assignment (Hamid et al., 2021). However, dopamine release supports different learning modalities in different compartments of the striatum (Lerner et al., 2015; Matsumoto and Hikosaka, 2009). Moreover, dopamine release occurs in wave-like patterns that propagate in the medial-lateral axis of the dorsal striatum, and the direction of wave propagation depends on behavioral context (Hamid et al., 2021). Specifically, tasks in which reward depended on animal performance promoted dopamine release beginning in the dorsomedial striatum, and tasks in which reward was independent of animal behavior promoted dopamine release beginning in the dorsolateral striatum. The authors (Hamid et al., 2021) performed computer simulations to illustrate the hypothesis that spatiotemporal gradients in dopamine release could emphasize reward credit assignment to specialized striatal subregions and recruit behavioral strategies appropriate for the task at hand. Temporal coincidence in the activity of multiple signals may be involved in functional segregation of the striatal subregions (Aosaki et al., 1994; Morris et al., 2004). A joint consideration of both spatio-temporal (Hamid et al., 2021) and modally (Lerner et al., 2015; Matsumoto and Hikosaka, 2009) mixed dopamine signals could further differentiate the function of different compartments of the striatum.

Predictions and Limitations

A key feature of learning is the transition from exploration to persistently performing a rewarding behavior. Here, we show in simulation how different learning modalities across the DMS and DLS could interact to contribute to this transition. The balance between exploration of unknown options and exploitation of known rewards is a ubiquitous challenge necessary for the survival of animals (Wilson et al., 2021). Even after learning new reward-feedback rules, animals sample alternate actions, and this alternative sampling is representative of continual exploration (Ebitz et al., 2019; Pisupati et al.,

2021). The neurophysiological basis for exploration and strategic control is an active topic of research with focus in the cortex (Cervera et al., 2020; Kaanders et al., 2021).

Our results here suggest a mechanism for the involvement of the DMS in the support of alternative sampling after a reward-feedback rule has been learned. Once behavior becomes driven primarily by stimulus-response association--in other words by the DLS in the lateral partition--cortico-striatal synaptic weights in the DMS are not required to maintain agent behavioral performance. The reward prediction error remains small as the agent persistently selects the rewarded action, and synaptic weights decay in the DMS. As these weights decay, the performance of outcome selection in the medial partition lags action selection in the lateral partition. On individual trials, the PFC—the cortical compartment associated with the medial partition—increasingly disagrees with the lateral partition—which determines behavior. What role, then does outcome selection in the medial partition play after extensive training?

These mistakes or lapses in the medial partition do increase the likelihood of alternative sampling in action selection but are not directly detectable through observation of behavior. Synaptic weight update in the DMS is contingent upon reward prediction error, so mistakes or lapses of outcome selection in the PFC do not result in learning unless the selection of an unrewarded action is realized in behavior. We formulate this observation as a key prediction of our computational model: neuronal activity in the PFC should occasionally and increasingly correspond to alternative unrewarded actions as an animal learns to persistently select a rewarded action. In trials when the PFC codes for an unrewarded action, that action may or may not be selected. The model predicts that instances of alternative sampling in outcome selection do not contribute to learning in the DMS unless that unrewarded action is reflected in behavior. In that case, the negative reward prediction error coded in dopamine release in the DMS should alter cortico-striatal weights to decrease the likelihood to select that unrewarded action. After a time, the strength of these synapses relax and the likelihood to select

that unrewarded action increases until it is sampled again. In this way, the PFC and DMS could monitor alternative reward contingencies in a constrained fashion and contribute to tonic exploration.

In initial learning, reward reversal tasks, and punishment tasks, we observed a discrepancy in steady-state performance between medial and lateral partitions. In sessions where a new stimulus-response association is learned, the PMC reliably selects the newly rewarded action. However, the PFC does not perform as well as the PMC. A tendency to sample previously rewarded actions—in other words tonic exploration—may indicate improved behavioral flexibility (Ebitz et al., 2019). Macaques that are more likely to sample previously rewarded actions after learning a new reward rule make fewer perseverative errors immediately following reward-rule changes (Ebitz et al., 2019). Following chronic exposure to cocaine, macaques are less likely to sample previously rewarded actions and are more likely to make perseverative errors immediately following a change in reward rule (Ebitz et al., 2019). We relate these results to our reward reversal $\mathrm{PFC}^{\dagger}$ simulations: in simulations with a misaligned PFC, agents exhibit a substantial increase in perseverative errors (Fig. 7A). In simulations of these behavioral sessions, outcome representation in the PFC did not correspond well to the two possible actions. As a technical consequence of PFC misalignment, the firing rate in the PFC that effectively determined outcome selection was low compared to control simulations, and thus the change in weights due to reward feedback was relatively slow. Since model behavior in early tasks was dependent on goal-directed learning, this manipulation compromised the ability of agents to overcome stimulus-response associations for previously rewarded actions. The comparative steady state performance of PFC and PMC in our results differed from (Ebitz et al., 2019). Steady state performance in this model is inextricably linked to the ability of medial partition to drive goal-directed behavior and the fidelity of outcome representation in PFC-DMS weights. The misalignment of the PFC compromised the ability of the medial partition to accumulate information about outcomes: the PFC-DMS D1 and D2 weights in these sessions did not differ from baseline as much as in control simulations. This phenomenon

restricted the performance of the PFC and increased its susceptibility to the loss of outcome representation in PFC-DMS synaptic weights via weight decay. The resulting poor performance in outcome selection contributed to increased selection of the previously rewarded action by the PMC. Inflexibility in this model manifests as a disruption in credit assignment to striatal actors so that new reward rules are slow to be discovered.

Our model implements the cortex in a limited fashion, and in future works, this aspect of the model could be extended to capture greater anatomical and physiological detail. In our previous models, stimulus-response associations were captured by Hebbian learning in cortico-cortical synaptic weights (Kim et al., 2017; Mulcahy et al., 2020). Plasticity in these projections was omitted in the present study to emphasize the role of DLS in the acquisition of stimulus-response associations and habit. However, cortical mechanisms for learning are critical for cognitive and executive flexibility (Floresco, 2013). The inclusion of cortical learning would permit action representations external to the striatum and allow us to investigate the cortical dopamine dynamics of learning tasks in healthy animals and in animals that suffer from the chronic use of addictive substances. Our present model does not distinguish between different regions of the prefrontal cortex. The prelimbic and infralimbic cortices appear to play opposing roles in learning (Amaya and Smith, 2018; Lipton et al., 2019). Disruption of infralimbic cortex prevents the expression of habit and facilitates goal-directed behavior (Coutureau and Killcross, 2003; Smith et al., 2012), and the interaction of the prelimbic and infralimbic regions is critical for the performance of set shifting tasks (Mukherjee and Caroni, 2018). Expanding this model to include the specific functions of these brain regions may accommodate aspects of behavior where learning resembles an immediate state change (e.g. attentional set shifting) as opposed to gradual changes over trials.

Methods

Neuronal simulations.

Neuronal populations are modeled by their firing rates, and these firing rates are defined by ordinary differential equations. The instantaneous firing rate A is modeled as

$$\tau \frac{dA}{dt} = \sigma(I) - A + N(t)$$

The time constant is set to 12.8 msec for the STN, 20 msec for the GPe, and 15 ms for all other neuronal populations (Pavlides et al., 2015). The properties of neuronal populations are determined by the connectivity of neural circuitry, which was manifest as differences in their input currents (I). The activation function was

$$\sigma(I) = \begin{cases} 0, & if\ I \leq 0 \\ \tanh(I), & if\ I > 0 \end{cases}$$

The term $N(t)$ is a noisy process with uniform distribution and amplitude 0.1. Numerical integration is performed using the Euler method with a timestep of 0.15 ms.

The neuronal connectivity is distributed within and across network partitions. Here, the indicator DXS represents one of either the dorsomedial striatum (DMS) or the dorsolateral striatum (DLS), and the indicator CTX represents one of either the prefrontal cortex (PFC) or the premotor cortex (PMC). Partitions are organized such that the DMS receives selective input from the PFC in the medial partition and the DLS receives selective input from the PMC in the lateral partition. With the exception of the STN, the definition of BG populations is identical between the medial and lateral partitions. Circuitry within the BG is organized to participate in cortico-BG-thalamo-cortical loops which compete via inhibition at the level of the cortex. These loops correspond to individual processing channels within the BG. Within the each of the medial and lateral partition, there are two competing loops that correspond to action #1 and action #2.

$$I_{D1} = \mathrm{g}_{CTX} w_{CTX-D}\ CTX$$

$$I_{D2} = \mathrm{g}_{CTX} w_{CTX-D2} CTX$$

$$I_{GPe} = dr_{GPe} - w_{D2-GP}\ D2 + w_{STN-GPe} STN$$

$$I_{STN} = I_{STN,DXS}$$

$$I_{GPi} = dr_{GPi} - w_{D1-GPi} D1 - w_{STN-GPi} STN$$

The variable for each neuronal population corresponds to its firing rate. Input currents were determined by the weights of their synaptic inputs (w)—which modified the firing rate of the presynaptic population,their excitatory drives (dr), and in some cases a gain (g). The input current of the STN differs between DMS and DLS partitions. The STN populations within the DLS receive hyperdirect pathway input from the PMC:

$$I_{STN,DMS} = dr_{STN} - w_{GPe-STN} GPe$$

$$I_{STN,DLS} = dr_{STN} - w_{GPe-STN} GPe + w_{HD} PMC$$

Mutual inhibition within the cortex (PFC-to-PFC inhibition and PMC-to-PMC inhibition) is tuned to force winner-takes-all choices. This outcome selection within the PFC is influenced by inhibitory inputs from the GPi_{DMS}. Action selection within in PMC is influenced by excitatory input from the PFC as well as inhibitory input from the GPi_{DLS}.

$$I_{PFC_m} = dr_{PFC_m} - w_{GPi-PFC} GPi_{DMS} - w_{PFC_n-PFC_m} PFC_n$$

$$I_{PMC_m} = dr_{PMC_m} + w_{PFC-PMC} PFC_m - w_{GPi-PMC} GPi_{DLS} - w_{PFM_n-PMC_m} PMC_n$$

Sessions with misaligned PFC

In $\mathrm{PFC}^{\dagger}$ sessions, we alter the connectivity to and from the PFC in order to simulate scenarios in which PFC activity was not well corelated with action selection. A pair of effective PFC ($eff\mathrm{PFC}$) firing rates were produced, which is a weighted average of the firing rate of the two PFC populations:

$$eff\mathrm{PFC}_m = 0.9\ PFC_m + 0.1\ PFC_n$$

Notice that the "wrong" channel PFC_n contributes 10% to $eff\mathrm{PFC}_m$. If each channel contributed equally (50%) to $eff\mathrm{PFC}_m$, then the PFC would be completely uncorrelated to action selection: $eff\mathrm{PFC}_m$ and $eff\mathrm{PFC}_n$ would each contain the same signal. Since the contribution 10% is one fifth of the way to equal contribution to $eff\mathrm{PFC}$, the PFC is said to be misaligned with action selection by 20%. In these sessions, $eff\mathrm{PFC}$ firing rates are used in place of PFC firing rates to compute the synaptic currents on D1 MSNs, D2 MSNs, and in PMC populations. Similarly, the inhibitory input onto PFC populations from the GPi are adjusted to be a combination of inputs from competing loops:

$$I_{PFC_m} = dr_{PFC_m} - w_{GPi-PFC}\left(0.9\ GPi_{DMS,m} + 0.1\ GPi_{DMS,n}\right) - w_{PFC_n-PFC_m}PFC_n$$

In this way, PFC activity is adjusted to carry less information about action selection.

Biophysical parameters are defined in Table 2.

Synaptic plasticity.

Dopaminergic input to the DMS codes the reward prediction error, which is the difference between reward feedback (R) and the expected reward ($ExpRew$).

$$SNc_{DMS} = R - ExpRew_j$$

$$ExpRew_{j+1} = \alpha R_j + (1-\alpha)ExpRew_j$$

The constant α takes value 0.15, and $ExpRew_j$ refers to the expected reward of the jth trial. For most simulations, the dopaminergic input to the DLS is identical to the expected reward. However, when the reward feedback is negative (indicating an aversive stimulus), the absolute value of the reward feedback is used to compute the rectified expectation ($RectExp$).

$$SNc_{DLS} = RectExp_j$$

$$RectExp_{j+1} = \alpha\left|R_j\right| + (1-\alpha)RectExp_j$$

The quantities SNc_{DMS} and SNc_{DLS} respectively represent the dopaminergic input to the DMS and DLS, which are utilized after each trial to updated cortico-striatal weights. Incremental updates to these weights are determined by the following expressions:

$$\Delta w_{PFC-D1,DMS,m} = \lambda_{DMS,D1} * SNc_{DMS} * PFC_m * D1_{DMS,m} - d * (w_0 - w_{PFC-D1,DMS,m})$$

$$\Delta w_{PFC-D2,DMS,m} = -\lambda_{DMS,D2} * SNc_{DMS} * PFC_m * D2_{DMS,m} - d * (w_0 - w_{PFC-D1,DMS,m})$$

$$\Delta w_{PMC-D1,DLS,m} = \lambda_{DLS,D1} * SNc_{DLS} * PMC_m * D1_{DLS,m} - d * (w_0 - w_{PMC-D1,DLS,m})$$

$$\Delta w_{PMC-D2,DLS,m} = -\lambda_{DLS,D2} * SNc_{DLS} * PMC_m * D2_{DLS,m} - d * (w_0 - w_{PMC-D1,DLS,m})$$

Weight updates were determined by a learning rate (λ) which was specific to the striatal population and a decay rate (d). Weight decay was parameterized by the resting weight (w_0). If an update reduced a weight to be negative, it was reset to be 0.

Parameter name	Parameter value
g_{PFC}	0.4
g_{PMC}	1
dr_{GPe}	1.6
w_{D2-GPe}	2
$w_{STN-GPe}$	0.4
dr_{STN}	0.8
$w_{GPe-STN}$	1
w_{HD}	0.3
dr_{GPi}	0.2
w_{D1-GPi}	1.4
$w_{STN-GPi}$	1.6
dr_{PFC}	1.5

$w_{GPi-PFC}$	1.8
$w_{PFC-PFC}$	1.6
$w_{PFC-PMC}$	0.1
$w_{GPi-PMC}$	1.8
$w_{PMC-PMC}$	1.6
$\lambda_{DMS,D1}$	0.05
$\lambda_{DMS,D2}$	-0.025
$\lambda_{DLS,D1}$	0.0025
$\lambda_{DLS,D1}$	0.00125
d	0.02
w_0	1

Table 2. Biophysical parameter values.

Behavioral task definitions.

Our computational model was challenged with several behavioral tasks. In a single trial, a neuronal simulation would produce the selection of an action, and based on the selected action, some reward feedback would be delivered. The reward feedback was used to update cortico-striatal weights. During the subsequent trial, the performance of action selection was determined by these newly updated weights. In this way, the model learns by incorporating reward feedback to alter the performance of action selection.

We produced 100 instances of this model, and each model instance was used to perform agent-based simulations. All agents were first trained with an Initial Learning session that was 200 trials long. During this session, the reward feedback for each trial was set to 1 for the selection of action #1 and 0 for the selection of action #2. Following the Initial Learning session, each agent was challenged with five additional behavioral tasks which had different reward feedback rules: Reward Devaluation, Reward

Reversal, Punishment, Reward Reversal $PFC^{\dagger}$, and Punishment $PFC^{\dagger}$ (see Table 3). In each of these sessions, the cortico-striatal weights were inherited from an Initial Learning session; in other words, agents were already trained to select action #1. In Reward Devaluation sessions, the reward feedback for action #1 was set to 0.2. In Reward Reversal sessions, the reward feedback was set to 0 for action #1 and 1 for action #2. In the Punishment session, the reward feedback was set to -0.5 for action #1 and 0 for action #2. The reward feedback for the Reward Reversal $PFC^{\dagger}$ and Punishment $PFC^{\dagger}$ sessions were identical to the Reward Reversal and Punishment sessions, but the PFC neuronal population connectivity was altered in the $PFC^{\dagger}$ sessions. Each of the five follow-up sessions contained 2000 trials.

Session Name	Action #1 feedback	Action #2 feedback	No. Trials
Initial Learning	1	0	200
Devaluation	0.2	0	2000
Reward Reversal	0	1	2000
Punishment	-0.5	0	2000
Reward Reversal $\mathrm{PFC}^{\dagger}$	0	1	2000
Punishment $\mathrm{PFC}^{\dagger}$	-0.5	0	2000

Table 3. Behavioral sessions were defined by the organization of reward feedback.

Evaluating model performance.

The group performance of agents in Reward Reversal and Punishment sessions was quantified using a normative model of evidence accumulation (Kilpatrick et al., 2019). Change points in agent performance were measured by detecting zeros of the log likelihood ratio of outcome selection (based on PFC activity) and action selection (based on PMC activity). We defined the log likelihood ratio for cortical selection to be

$$y_n = \log\frac{P_1(w_n)}{P_2(w_n)} + \log\frac{(1-h)\exp(y_{n-1}) + h}{h\exp(y_{n-1}) + (1-h)}$$

The hazard coefficient h was fixed to $1/201$; the reward feedback rules changed exactly once after the 200 trials of the Initial Learning session. When analyzing cortical activity, P_1 and P_2 were the likelihood of the agent selecting a particular outcome or action. This selection is determined by cortico-striatal

weights (w_n for the $\mathrm{n}th$ trial) as well as noise added to the initial conditions and in each step of numerical integration for that trial. We measured P_1 by caching the cortico-striatal weights for every trial in each session that an agent performs, repeating each trial 1000 times with different seeds for the pseudo random number generator, and recording the number of times outcome #1 was selected by the PFC or action #1 was selected by the PMC. We defined $P_2 = 1 - P_1$. The change point in PFC or PMC performance was determined by the trial number of the first zero-crossing of the log likelihood ratio y.

The difference in cortical selection performance of Reward Reversal and Punishment sessions to Reward Reversal $\mathrm{PFC}^{\dagger}$ and Punishment $\mathrm{PFC}^{\dagger}$ sessions was quantified by comparing the changepoints in PFC and PMC activity. Since changepoints indicated integer trial numbers, these comparisons were performed using the Mann-Whitney U test with an alpha level of 0.05.

We evaluated performance of agents in the PFC and the PMC at the end of each behavioral session. Steady state performance of an individual agent was determined by P_1 and P_2 for the final trial of that session. We quantified the group difference between PFC and PMC performance between agents using the Mann-Whitney U test with an alpha level of 0.05. Only P_1 was used for these comparisons since P_2 was exactly determined by P_1.

Software

Simulations were performed in Python using the JAX package (Bradbury et al., 2018). Simulation results were analyzed using NumPy and SciPy (Harris et al., 2020; Virtanen et al., 2020) . Statistics were computed using the wilcox.test function in the R computing environment. Figures were created with Matplotlib (Hunter, 2007), Adobe Illustrator, and Microsoft PowerPoint. All models and simulations will be made free and open source following publication of the peer reviewed manuscript.

Acknowledgements

This work was supported by AA023786, P60-AA007611, and 5T32AA007462. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

Bibliography

Amaya, K.A., and Smith, K.S. (2018). Neurobiology of habit formation. Current Opinion in Behavioral Sciences *20*, 145–152.

Aosaki, T., Tsubokawa, H., Ishida, A., Watanabe, K., Graybiel, A., and Kimura, M. (1994). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. J Neurosci *14*, 3969–3984.

Balleine, B.W., and O'Doherty, J.P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology *35*, 48–69.

Bariselli, S., Fobbs, W.C., Creed, M.C., and Kravitz, A.V. (2019). A competitive model for striatal action selection. Brain Res *1713*, 70–79.

Barker, J.M., Corbit, L.H., Robinson, D.L., Gremel, C.M., Gonzales, R.A., and Chandler, L.J. (2015). Corticostriatal circuitry and habitual ethanol seeking. Alcohol *49*, 817–824.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. (2018). JAX: composable transformations of Python+NumPy programs.

Cervera, R.L., Wang, M.Z., and Hayden, B.Y. (2020). Systems neuroscience of curiosity. Current Opinion in Behavioral Sciences *35*, 48–55.

Corbit, L.H., and Janak, P.H. (2016). Habitual Alcohol Seeking: Neural Bases and Possible Relations to Alcohol Use Disorders. Alcohol Clin Exp Res *40*, 1380–1389.

Corbit, L.H., Nie, H., and Janak, P.H. (2012). Habitual Alcohol Seeking: Time Course and the Contribution of Subregions of the Dorsal Striatum. Biological Psychiatry *72*, 389–395.

Coutureau, E., and Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. Behav Brain Res *146*, 167–174.

Crews, F.T., and Boettiger, C.A. (2009). Impulsivity, frontal lobes and risk for addiction. Pharmacol Biochem Behav *93*, 237–247.

Cui, G., Jun, S.B., Jin, X., Pham, M.D., Vogel, S.S., Lovinger, D.M., and Costa, R.M. (2013). Concurrent Activation of Striatal Direct and Indirect Pathways During Action Initiation. Nature *494*, 238–242.

De Falco, E., White, S.M., Morningstar, M.D., Ma, B., Nkurunziza, L.T., Ahmed-Dilibe, A., Wellman, C.L., and Lapish, C.C. (2021). Impaired cognitive flexibility and heightened urgency are associated with increased alcohol consumption in rodent models of excessive drinking. Addict Biol e13004.

Donahue, C.H., and Lee, D. (2015). Dynamic Routing of Task-relevant Signals for Decision Making in Dorsolateral Prefrontal Cortex. Nat Neurosci *18*, 295–301.

Durstewitz, D., Seamans, J.K., and Sejnowski, T.J. (2000). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. J Neurophysiol *83*, 1733–1750.

Ebitz, R.B., Sleezer, B.J., Jedema, H.P., Bradberry, C.W., and Hayden, B.Y. (2019). Tonic exploration governs both flexibility and lapses. PLOS Computational Biology *15*, e1007475.

Floresco, S.B. (2013). Prefrontal dopamine and behavioral flexibility: shifting from an "inverted-U" toward a family of functions. Front Neurosci *7*.

Frank, M.J. (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.
Graybiel, A.M. (2008). Habits, rituals, and the evaluative brain. Annu Rev Neurosci *31*, 359–387.
Hamid, A.A., Frank, M.J., and Moore, C.I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. Cell.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.
Hart, G., Bradfield, L.A., and Balleine, B.W. (2018a). Prefrontal Corticostriatal Disconnection Blocks the Acquisition of Goal-Directed Action. J Neurosci *38*, 1311–1322.
Hart, G., Bradfield, L.A., Fok, S.Y., Chieng, B., and Balleine, B.W. (2018b). The Bilateral Prefronto-striatal Pathway Is Necessary for Learning New Goal-Directed Actions. Curr Biol *28*, 2218-2229.e7.
Hunnicutt, B.J., Jongbloets, B.C., Birdsong, W.T., Gertz, K.J., Zhong, H., and Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization. ELife *5*, e19103.
Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science Engineering *9*, 90–95.
Jenni, N.L., Larkin, J.D., and Floresco, S.B. (2017). Prefrontal Dopamine D1 and D2 Receptors Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar Circuits. J Neurosci *37*, 6200–6213.
Kaanders, P., Juechems, K., O'Reilly, J., and Hunt, L. (2021). Dissociable mechanisms of information sampling in prefrontal cortex and the dopaminergic system. Current Opinion in Behavioral Sciences *41*, 63–70.
Kennerley, S.W., Dahmubed, A.F., Lara, A.H., and Wallis, J.D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. J Cogn Neurosci *21*, 1162–1178.
Kesner, R.P., and Churchwell, J.C. (2011). An analysis of rat prefrontal cortex in mediating executive function. Neurobiol Learn Mem *96*, 417–431.
Kilpatrick, Z.P., Holmes, W.R., Eissa, T.L., and Josić, K. (2019). Optimal models of decision-making in dynamic environments. Current Opinion in Neurobiology *58*, 54–60.
Kim, T., Hamade, K.C., Todorov, D., Barnett, W.H., Capps, R.A., Latash, E.M., Markin, S.N., Rybak, I.A., and Molkov, Y.I. (2017). Reward Based Motor Adaptation Mediated by Basal Ganglia. Front. Comput. Neurosci. *11*.
Lak, A., Stauffer, W.R., and Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. Proc Natl Acad Sci U S A *111*, 2343–2348.
Laubach, M., Amarante, L.M., Swanson, K., and White, S.R. (2018). What, If Anything, Is Rodent Prefrontal Cortex? ENeuro *5*.
Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R., et al. (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell *162*, 635–647.

Linsenbardt, D.N., Timme, N.M., and Lapish, C.C. (2019). Encoding of the Intent to Drink Alcohol by the Prefrontal Cortex Is Blunted in Rats with a Family History of Excessive Drinking. ENeuro *6*.
Lipton, D.M., Gonzales, B.J., and Citri, A. (2019). Dorsal Striatal Circuits for Habits, Compulsions and Addictions. Front Syst Neurosci *13*, 28.
Lüscher, C., and Janak, P.H. (2021). Consolidating the Circuit Model for Addiction. Annual Review of Neuroscience *44*, null.
Lüscher, C., Robbins, T.W., and Everitt, B.J. (2020). The transition to compulsion in addiction. Nat Rev Neurosci *21*, 247–263.
Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.
Miller, K.J., Shenhav, A., and Ludvig, E.A. (2019). Habits without values. Psychol Rev *126*, 292–311.
Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron *43*, 133–143.
Mukherjee, A., and Caroni, P. (2018). Infralimbic cortex is required for learning alternatives to prelimbic promoted associations through reciprocal connectivity. Nat Commun *9*, 2727.
Mulcahy, G., Atwood, B., and Kuznetsov, A. (2020). Basal ganglia role in learning rewarded actions and executing previously learned choices: Healthy and diseased states. PLOS ONE *15*, e0228081.
Murray, J.E., Belin, D., and Everitt, B.J. (2012). Double dissociation of the dorsomedial and dorsolateral striatal control over the acquisition and performance of cocaine seeking. Neuropsychopharmacology *37*, 2456–2466.
Myrick, H., Anton, R.F., Li, X., Henderson, S., Drobes, D., Voronin, K., and George, M.S. (2004). Differential brain activity in alcoholics and social drinkers to alcohol cues: relationship to craving. Neuropsychopharmacology *29*, 393–402.
Ostlund, S.B., and Balleine, B.W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J Neurosci *25*, 7763–7770.
Ott, T., and Nieder, A. (2019). Dopamine and Cognitive Control in Prefrontal Cortex. Trends in Cognitive Sciences *23*, 213–234.
Pan, W.X., Mao, T., and Dudman, J.T. (2010). Inputs to the dorsal striatum of the mouse reflect the parallel circuit architecture of the forebrain. Front Neuroanat *4*, 147.
Pavlides, A., Hogan, S.J., and Bogacz, R. (2015). Computational Models Describing Possible Mechanisms for Generation of Excessive Beta Oscillations in Parkinson's Disease. PLoS Comput Biol *11*, e1004609.
Perry, J.L., Joseph, J.E., Jiang, Y., Zimmerman, R.S., Kelly, T.H., Darna, M., Huettl, P., Dwoskin, L.P., and Bardo, M.T. (2011). Prefrontal cortex and drug abuse vulnerability: translation to prevention and treatment interventions. Brain Res Rev *65*, 124–149.
Peters, A.J., Fabre, J.M.J., Steinmetz, N.A., Harris, K.D., and Carandini, M. (2021). Striatal activity topographically reflects cortical activity. Nature 1–6.
Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., and Churchland, A.K. (2021). Lapses in perceptual decisions reflect exploration. Elife *10*.

Puig, M.V., Rose, J., Schmidt, R., and Freund, N. (2014). Dopamine modulation of learning and memory in the prefrontal cortex: insights from studies in primates, rodents, and birds. Front. Neural Circuits *8*.
Rossi, M.A., and Yin, H.H. (2012). Methods for studying habitual behavior in mice. Curr Protoc Neurosci *Chapter 8*, Unit 8.29.
Schacht, J.P., Anton, R.F., and Myrick, H. (2013). Functional neuroimaging studies of alcohol cue reactivity: a quantitative meta-analysis and systematic review. Addict Biol *18*, 121–133.
Schwabe, L., and Wolf, O.T. (2011). Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. Behav Brain Res *219*, 321–328.
Smith, K.S., Virkud, A., Deisseroth, K., and Graybiel, A.M. (2012). Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. Proc Natl Acad Sci U S A *109*, 18932–18937.
Tecuapetla, F., Jin, X., Lima, S.Q., and Costa, R.M. (2016). Complementary Contributions of Striatal Projection Pathways to Action Initiation and Execution. Cell *166*, 703–715.
Timme, N.M., Ma, B., Linsenbardt, D., Cornwell, E., Galbari, T., and Lapish, C. (2021). Compulsive drinking is associated with neural activity patterns reflecting diminished behavioral control and enhanced seeking representations in dorsal medial prefrontal cortex. BioRxiv 2021.03.15.435169.
Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. Science *307*, 1642–1645.
Trantham-Davidson, H., and Chandler, L.J. (2015). Alcohol-induced alterations in dopamine modulation of prefrontal activity. Alcohol *49*, 773–779.
Trantham-Davidson, H., Burnett, E.J., Gass, J.T., Lopez, M.F., Mulholland, P.J., Centanni, S.W., Floresco, S.B., and Chandler, L.J. (2014). Chronic alcohol disrupts dopamine receptor activity and the cognitive function of the medial prefrontal cortex. J Neurosci *34*, 3706–3718.
Tran-Tu-Yen, D.A.S., Marchand, A.R., Pape, J.-R., Di Scala, G., and Coutureau, E. (2009). Transient role of the rat prelimbic cortex in goal-directed behaviour. Eur J Neurosci *30*, 464–471.
Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S., and Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. Nat Commun *7*, 12554.
Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M.M., Uchida, N., and Watabe-Uchida, M. (2020). Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. ELife *9*, e62390.
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261–272.
Volkow, N.D., Wang, G.J., Fowler, J.S., Logan, J., Hitzemann, R., Ding, Y.S., Pappas, N., Shea, C., and Piscani, K. (1996). Decreases in dopamine receptors but not in dopamine transporters in alcoholics. Alcohol Clin Exp Res *20*, 1594–1598.
Volkow, N.D., Wang, G.-J., Maynard, L., Fowler, J.S., Jayne, B., Telang, F., Logan, J., Ding, Y.-S., Gatley, S.J., Hitzemann, R., et al. (2002). Effects of alcohol detoxification on dopamine D2 receptors in alcoholics: a preliminary study. Psychiatry Res *116*, 163–172.
Wilson, R.C., Bonawitz, E., Costa, V.D., and Ebitz, R.B. (2021). Balancing exploration and exploitation with information and randomization. Curr Opin Behav Sci *38*, 49–56.

Yin, H.H., and Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. Nature Reviews Neuroscience *7*, 464–476.
Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci *19*, 181–189.
Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci *22*, 513–523.
Yin, H.H., Mulcare, S.P., Hilário, M.R.F., Clouse, E., Holloway, T., Davis, M.I., Hansson, A.C., Lovinger, D.M., and Costa, R.M. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. Nature Neuroscience *12*, 333–341.

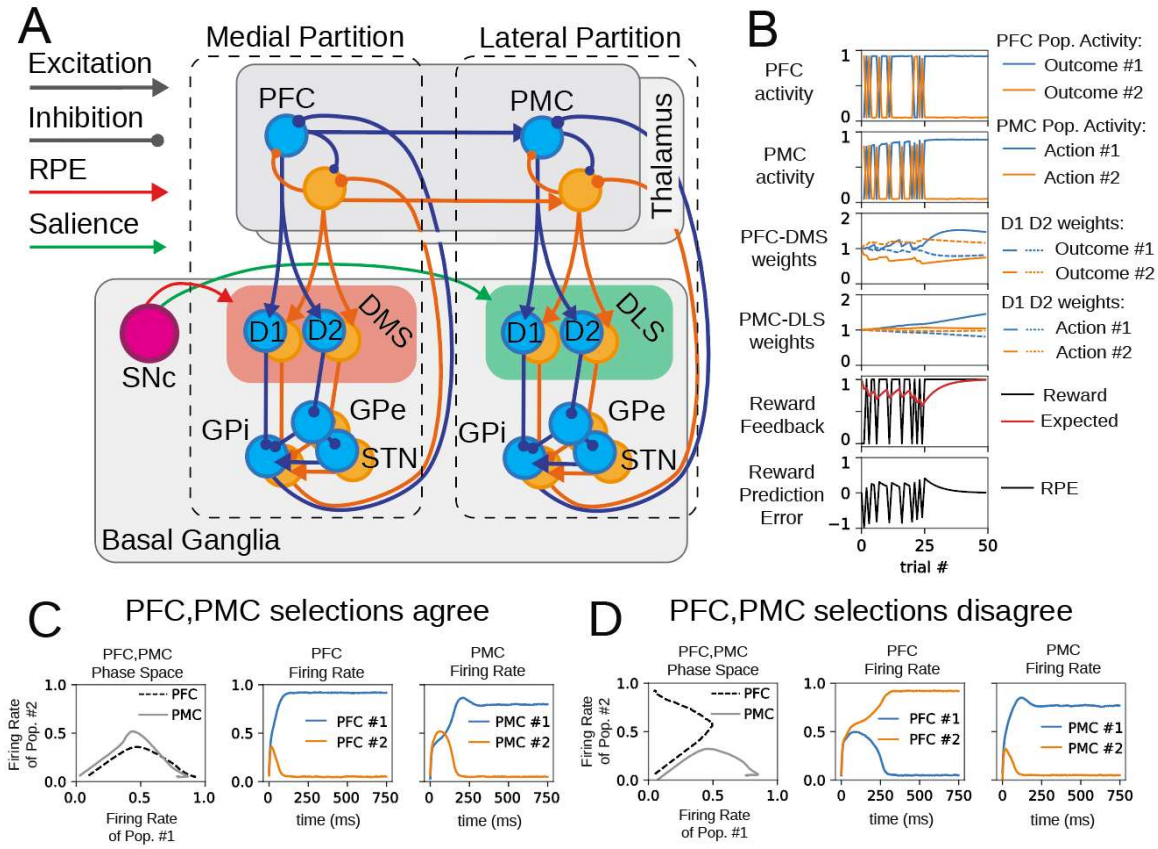


Figure 1. A model of bimodal and concurrent learning in the basal ganglia.

(A) Cortico-basal ganglia-thalamo-cortical loops are distributed between the medial and lateral partitions of the model and are organized into distinct channels that represent individual actions (color-coded as blue and orange). The medial partition incorporates the prefrontal cortex (PFC) and the portion of the basal ganglia that includes the dorsomedial striatum (DMS). The lateral partition incorporates the premotor cortex (PMC) and the portion of the basal ganglia that includes the dorsolateral striatum (DLS). The DMS is coded red to indicate that dopamine release in this region codes for reward prediction error. The DLS is coded green to indicate that dopamine release in this region codes for salience. Within each compartment of the basal ganglia, each of the two channels processes cortical input via the striatonigral (direct) pathway and striatopallidal (indirect) pathway which converge on a combined node representing the substantia nigra pars reticulata and the globus pallidus internal (SNr/GPi). D1-expressing medium spiny neurons (MSNs) directly inhibit the SNr/GPi (direct pathway). D2-expressing MSNs project to the indirect pathway, which includes the globus pallidus external (GPe) and the subthalamic nucleus (STN). Within each partition, cortical excitation of MSNs is modulated by dopaminergic projections from the substantia nigra compacta (SNc). These dopamine signals encode different quantities within the DMS and the DLS and thus induce different modes of learning in each partition. In the DMS, dopamine modulates cortico-striatal synaptic weights based on the reward prediction error (RPE), and in the DLS, dopamine modulates cortico-striatal weights based on contextual salience.

(B) Sample model activity during a behavioral session includes end-trial cortical activity, cortico-striatal synaptic weights, and reward feedback metrics.

(C,D) Cortical partitions (PFC and PMC) perform independent outcome and action selection. PFC and PMC selection may agree (C) or disagree (D).

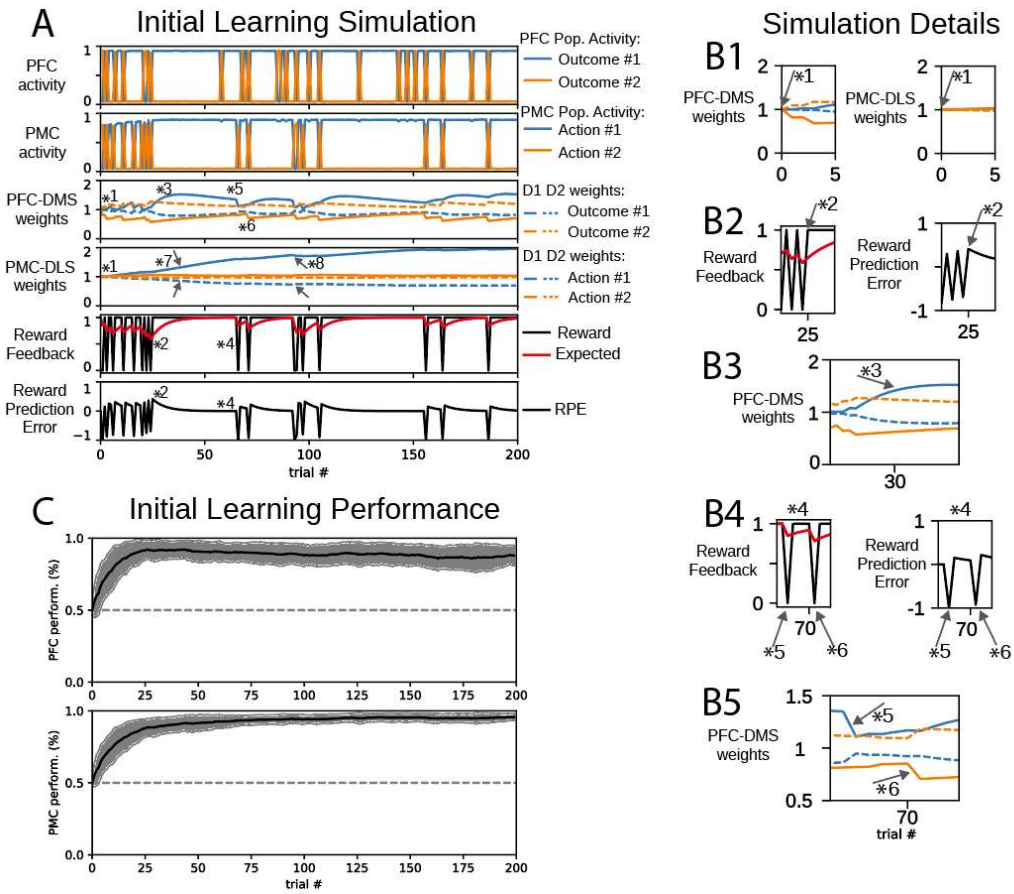


Figure 2.

(A) Cortical activity, cortico-striatal synaptic weights, and reward feedback metrics during initial learning behavioral session. In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond to action #1 and orange traces correspond to action #2.

(B) Insets correspond to trials indicated by markers 1-6 in part (A).

(B1) Initial D1, D2 weights were not biased to select either action. (B2) When the reward was greater than the expected reward, the reward prediction error (RPE) was positive. (B3) The PFC-DMS cortico-striatal synaptic weights evolve to promote the selection of outcome #1. (B4) When the expected reward is high and no reward is delivered, the reward prediction error becomes negative. (B5) Marker *5: selection of outcome #1 is de-emphasized when PFC selects outcome #1 and PMC selects the unrewarded action #2. Marker *6: selection of outcome #2 is de-emphasized when PFC selects outcome #2 and PMC selects the unrewarded action #2.

(C) Learning performance of 100 agents. Performance is defined as the likelihood for an agent to selection action #1.

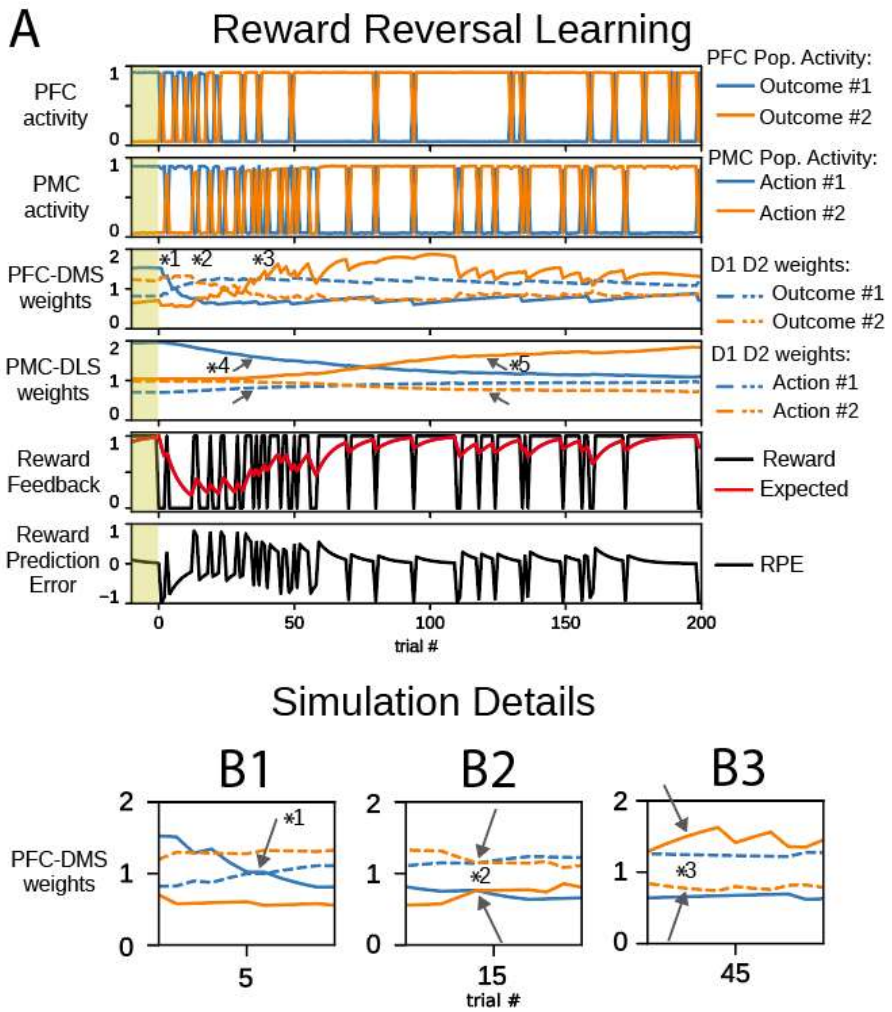


Figure 3. An agent learns to select action #2 in a reward reversal behavioral task.
(A) In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond to action #1 and orange traces correspond to action #2. The yellow shaded trials before trial 0 indicate trials at the end of the initial learning session.
(B) Insets correspond to trials indicated by markers in part (A).
(B1) PFC-DMS cortico-striatal synaptic weights respond to the change in reward feedback rule to stop selecting outcome #1. (B2) The PFC-DMS weights for both channels emphasize the indirect pathway; this configuration promotes exploration in the PFC. (B3) PFC-DMS cortico-striatal synaptic weights evolve to promote the selection of outcome #2.

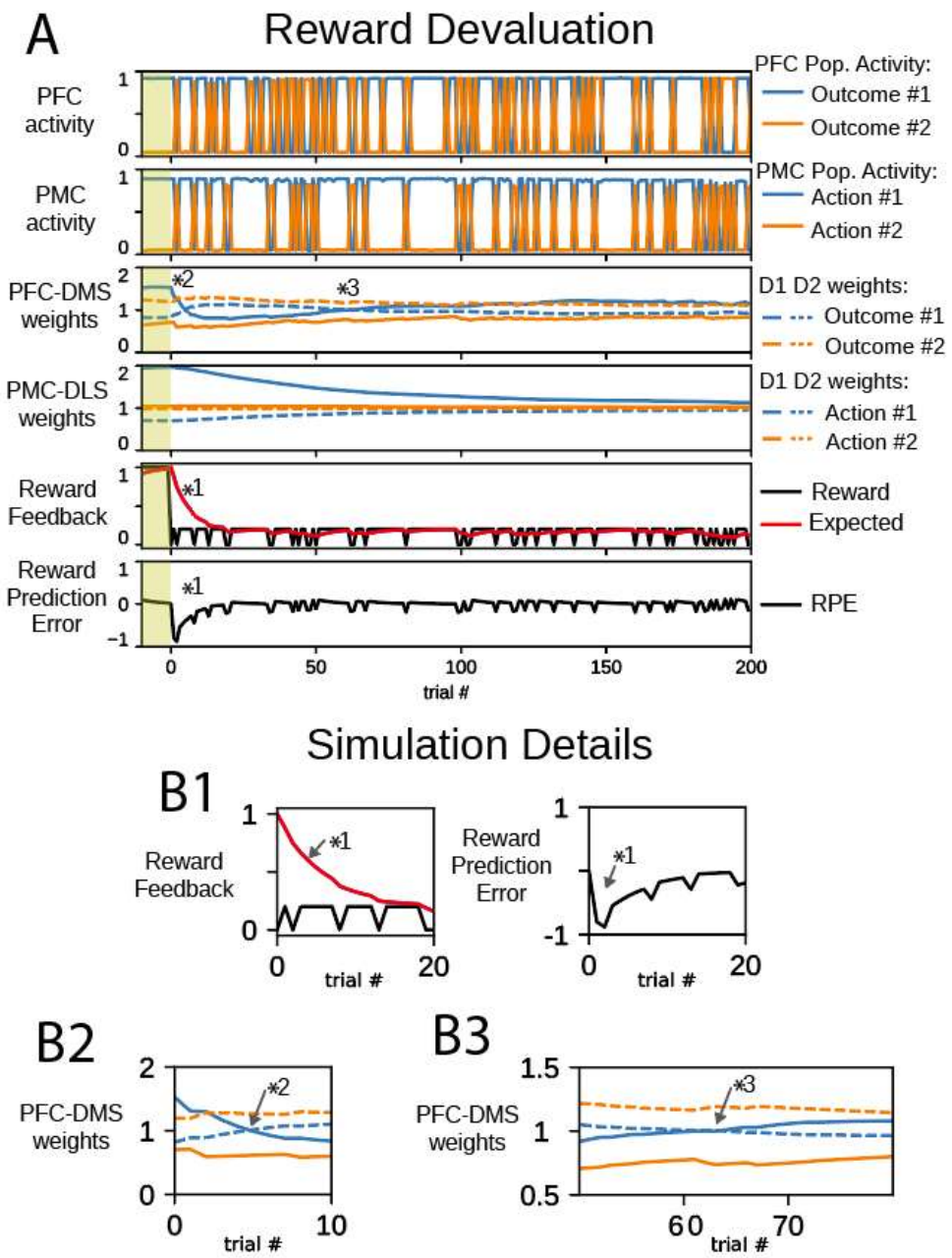


Figure 4. The magnitude of reward feedback is reduced in a reward devaluation behavioral task. The selection of action #1 persists but becomes less likely.
(A) In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond to action #1 and orange traces correspond to action #2. The yellow shaded trials before trial 0 indicate trials at the end of the initial learning session.
(B) Insets correspond to trials indicated by markers 1-3 in part (A). (B1) Following reward devaluation, the reward feedback is less than the expected reward. Even though, reward feedback continues to be mostly positive, the reward prediction error (RPE) is negative. (B2) The cortico-striatal synaptic weights for outcome #1 reflect the negative reward prediction error, and the indirect pathways for both outcomes are emphasized. (B3) Since action #1 is rewarded, the cortico-striatal synaptic weights for outcome #1 eventually recover to emphasize the direct pathway over the indirect pathway.

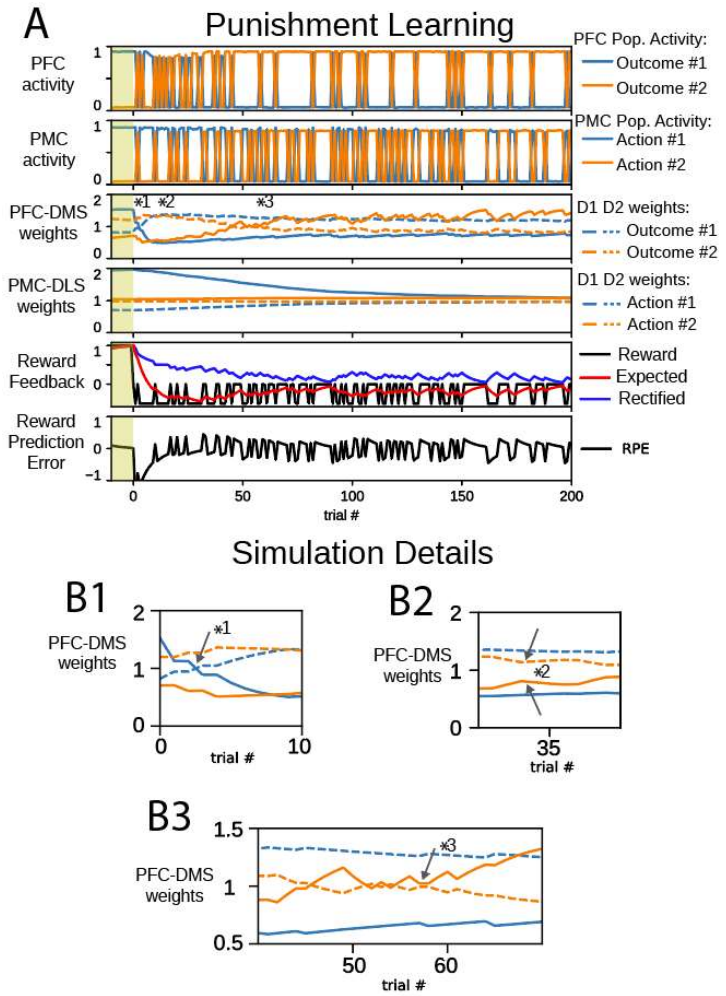


Figure 5. An agent learns to select action #2 in a punishment learning behavioral task.
Cortical activity, cortico-striatal synaptic weights, and reward feedback metrics evolve to overcome habit to select action #1. The yellow shaded trials before trial 0 indicate trials at the end of the initial learning session.
(A) In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond to action #1 and orange traces correspond to action #2.
(B) Insets correspond to trials indicated by markers 1-3 in part (A).
(B1) PFC-DMS cortico-striatal synaptic weights for outcome #1 respond to the negative reward prediction error and evolve to emphasize the indirect pathway. (B2) During exploration, the PFC-DMS channels for both outcomes emphasize the indirect pathway. (B3) In the channel for outcome #2, the direct pathway becomes emphasized to promote the selection of the unpunished choice.

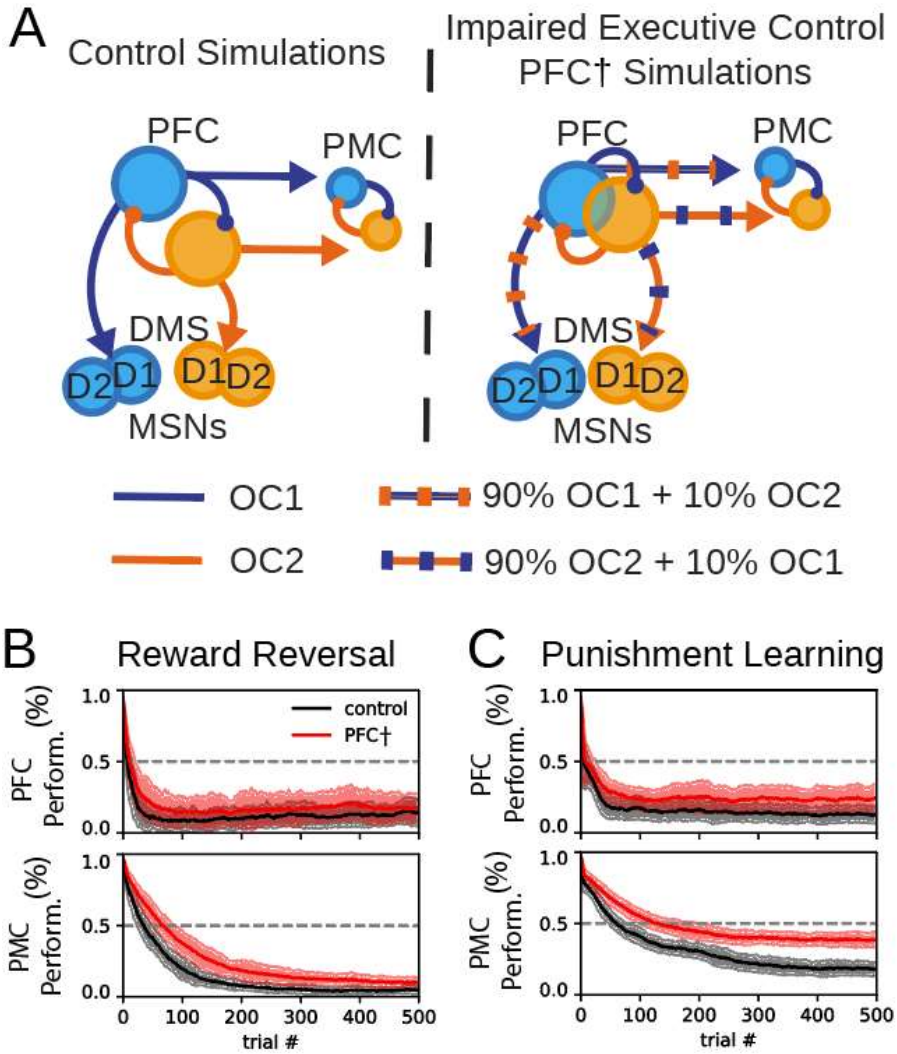


Figure 6. Impairment of PFC decreases learning performance.
(A). Illustration of impairment of PFC coding. The projections of the PFC into the DMS and the PMC contained mixed signals in $\mathrm{PFC}^{\dagger}$ behavioral sessions. OC1 and OC2 correspond to projections that represent the channels for outcome #1 and outcome #2.
(B-C) Performance is defined as the likelihood for agents (N=100) to selection action #1; here, agents learn to select action #2.
(B) Progression of outcome selection in the PFC and action selection in the PMC during a reward reversal task (black) and a reward reversal task with impaired PFC (red).
(C) Progression of outcome selection in the PFC and action selection in the PMC during a punishment learning task (black) and a punishment learning task with impaired PFC (red).

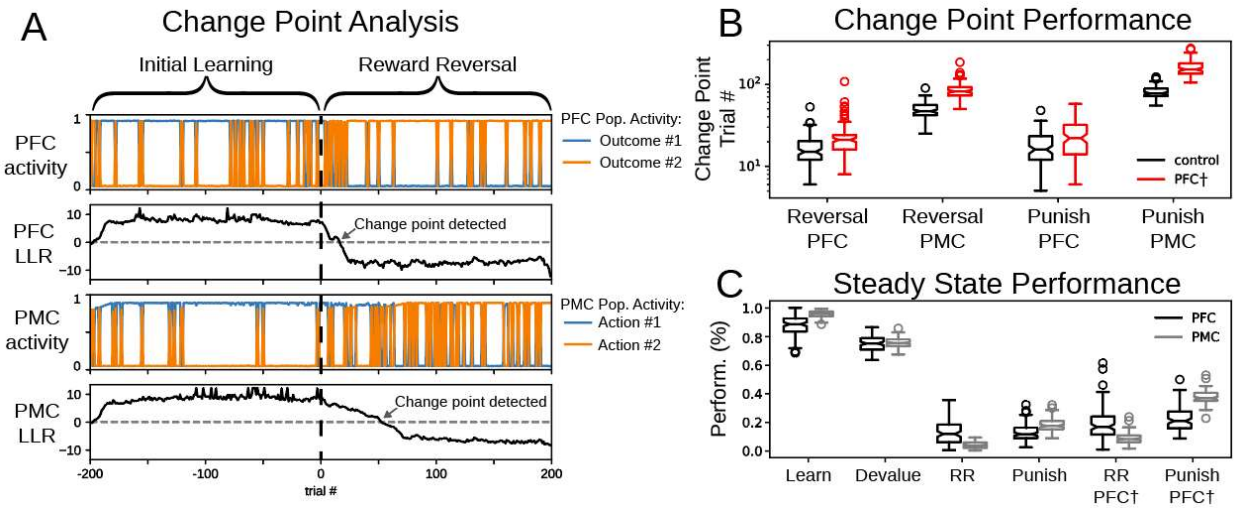


Figure 7. Analysis of model output across behavioral tasks

(A) Example model activity of an initial learning session and the following reward reversal session with change point analysis. Trial number here is relative to the beginning of the reward reversal session.

(B) Change point analysis indicates the trial at which an ideal observer detects a change in cortical selection from action #1 to action #2. We compared change point performance in control simulations (black) and sessions with impaired PFC (red) (N=100 agents). Box and whisker plots with median depicted.

(C) Performance of agents at the end of each behavioral task showing the likelihood that agents select action #1). We compared PFC performance (black) to PMC performance (grey) (N=100 agents). Box and whisker plots with median depicted.