Distinct cortico-striatal 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 the role of this circuitry in the emergence of inflexible behaviors such as those observed in addiction. Specifically, we develop a computational model of cortico-striatal interactions that performs concurrent goal-directed and habit 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 represents outcomes, and the DLS receives input from the premotor cortex which determines action selection. Following an initial learning of a two-alternative forced choice task, we subjected the model to reversal learning, reward devaluation, and learning a punished outcome. 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 impairment of executive control, which was implemented as poor outcome representation in the PFC. The degraded executive control reduced the efficacy of goal-directed learning, 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 circuits of the dorsal striatum are dynamically engaged to control behavior and how the impairment of executive control by the PFC enhances inflexible behavior.

Author Summary

Substance abuse changes an animal's brain so that it is not able to appropriately respond to new environmental cues such as negative consequences for drug use. It is thought that the transition from adaptive to inflexible responding involves the inappropriate engagement of neural circuitry responsible for habit. In this study, we model neural circuitry that is critical for adaptive responding to changes in reward feedback to animal behavior. We extended previous computational models of the basal ganglia to incorporate detailed biophysical representations of the dorsomedial striatum, which is critical for goal-directed behavior, and the dorsolateral striatum, which is critical for habitual behavior. The model exhibited learning in simulations of reward reversal, devaluation, and punished outcome tasks. Recent work from our group has quantified the impairment of executive control in cortical activity of compulsively drinking rats. Here, we challenged the model in additional simulations by implementing the impairment of executive control as a reduction in the fidelity of prefrontal cortical outcome representations. Our results illustrate how impaired executive control reduces the performance of goal-directed learning and, hence, reduces the model's ability to overcome an established habitual response.

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, and this transition involves 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 function of this circuitry from enabling the automaticity of routine tasks 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 impairment of executive control, such as addiction.

The dorsomedial striatum (DMS) and dorsolateral striatum (DLS) are basal ganglia structures that have been implicated in the pathophysiology of inflexible behavior and drug addiction (Corbit and Janak, 2016; Lipton et al., 2019). In early training, drug seeking behavior is disrupted by inactivation of the DMS but is insensitive to inactivation of the DLS (Corbit et al., 2012; Murray et al., 2012). A transition occurs after extensive training, where drug seeking is disrupted by inactivation of the DLS and is insensitive to inactivation of the DMS. While these data clearly outline distinct temporal roles of the medial and lateral compartments of the dorsal striatum in the control of behavior, how this transition occurs is unclear.

The distinct roles of the DMS and DLS in inflexible behavior are likely derived from differences in their computational properties. The DMS and DLS are, respectively, critical for goal-directed and habitual behavior (Schwabe and Wolf, 2011; Yin and Knowlton, 2006). The inactivation of the DMS decreases sensitivity to reward devaluation (Yin et al., 2005), and lesions of the DLS increases sensitivity to reward devaluation and abolishes habitual seeking in the absence of a reward (Yin et al., 2004). A potential mechanism for these differences was articulated by Yin et al. (2009) that found plasticity in the DLS occurs relatively slowly compared to DMS as training progresses.

The differences in the computational functions of the DMS and DLS are also supported by differences in dopamine efflux and cortical inputs. 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

82 where the medial portion of the striatum is more likely to receive input from the prefrontal cortex (PFC),
83 and the lateral portion more likely to receive input from somatosensory and motor regions (Hunnicutt et
84 al., 2016; Pan et al., 2010; Peters et al., 2021).

85 The input from PFC to the DMS is of particular interest. The PFC is important for cognition,
86 executive function and goal-directed behavior (Balleine and O'Doherty, 2010; Barker et al., 2015;
87 Donahue and Lee, 2015; Hart et al., 2018a, 2018b; Kennerley et al., 2009; Kesner and Churchwell, 2011;
88 Laubach et al., 2018; Ostlund and Balleine, 2005; Perry et al., 2011; Tran-Tu-Yen et al., 2009; Tsutsui et
89 al., 2016).. Neuronal activity in the PFC contextualizes sensory inputs with respect to cognitive state and
90 recent reward feedback (Mante et al., 2013; Niv, 2019; Redish, 2016), and these outcome
91 representations are necessary for goal-directed learning (Lipton et al., 2019). Critically, these regions are
92 impaired following prolonged alcohol use (Schacht et al., 2013) and this impairment corresponds to
93 increased responding for alcohol (Crews and Boettiger, 2009; Myrick et al., 2004). In recent work, we
94 investigated how the computational properties of the medial PFC are altered in a rodent model of
95 excessive alcohol consumption and found impairments in neural signatures of the intent to drink and
96 seeking behavior (Linsenbardt et al., 2019; Timme et al., 2021). Taken in combination with observations
97 about the DMS and DLS, we leverage these results to illustrate a hypothesis for how impairment in the
98 PFC contributes to the emergence of inflexible behavior.

99 Here we present a new computational model of cortico-striatal learning that incorporates goal-
100 directed learning in the DMS and habit learning in the DLS. We derive this implementation from a theory
101 of reinforcement learning based on dopamine-mediated plasticity of cortico-striatal projections to
102 medium spiny neurons (MSNs) (Graybiel, 2008). Dopamine induces long-term potentiation and long-
103 term depression in D1 and D2 receptor-expressing MSNs respectively, and these changes are
104 hypothesized to configure the basal ganglia to selectively disinhibit thalamocortical relay neurons in the
105 context of ongoing behavioral tasks (Frank, 2005). This hypothesis has been incorporated in

106 computational models that include reward-based learning that is based on RPE in the striatum (Frank,
107 2005; Kim et al., 2017; Mulcahy et al., 2020). Recent computational investigations of the basal ganglia
108 have accounted for stimulus-response learning as a feature of cortical plasticity (Baladron and Hamker,
109 2020; Kim et al., 2017; Mulcahy et al., 2020), as a feature of distinct dopaminergic reward coding in the
110 DLS (Bogacz, 2020; Miller et al., 2019; Tsutsui-Kimura et al., 2020), or in the context of the spatial
111 distribution of dopamine release in the medial-lateral axis of the dorsal striatum (Hamid et al., 2021). To
112 understand the mechanism of how behaviors transition from goal-directed to inflexible, a computational
113 model is required that is capable of articulating how changes in cortico-striatal plasticity support both
114 goal-directed learning in the DMS and stimulus-response learning in the DLS. Based on four key
115 assumptions (Table 1), we extend previous models that incorporate stimulus-response driven behavior
116 to elucidate the roles of the DMS and the DLS in adaptable behavior and how the circuitry responsible
117 for habitual responding could exhibit pathological primacy during inflexible behavior.

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

128

Experimental observation	**Citation**

DMS and DLS receive separate specific dopaminergic input	(Lerner et al., 2015; Matsumoto and Hikosaka, 2009)
DMS and DLS receive distinct cortical projections	(Hunnicutt et al., 2016; Pan et al., 2010; Peters et al., 2021)
Plasticity in the DLS is slower than plasticity in the DMS	(Yin et al., 2009)
Impairment of executive control attributed to poor PFC coding for outcomes	(Linsenbardt et al., 2019; Timme et al., 2021)

Table 1. Key additions that extend the model from prior works.

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 choice-specific channels (Frank, 2005; Kim et al., 2017; Mulcahy et al., 2020) that compete to perform reward-based learning across partitions of the dorsal striatum (Lerner et al., 2015; Matsumoto and Hikosaka, 2009). Choice-specific channels in this model correspond to representations of outcomes or actions in the neural activity of the cortex and basal ganglia. These channels are supported by the observation that MSN activity is organized into local ensembles of cells that encode outcome and action space (Klaus et al., 2017).

The model can, broadly, be separated into two partitions: The medial partition incorporates the PFC and the DMS, and the lateral partition incorporates the pre-motor cortex (PMC) and the DLS. Each partition also contains a direct and indirect pathway for each of the two behavioral choices that are presented to the agent. Each partition engages the machinery of the BG to utilize input from cortex, and the processing of this input modulates 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 simulated successive forced-choice trials. These forced choices are evaluated in both the medial and lateral cortico-striatal

loops. The neural activity in these two partitions disambiguates the neural representations of actions and the neural representations of outcomes that are contingent upon actions.

In this model, we distinguish between cortical inputs to the DMS and DLS, these inputs are the distinct cortical representations required to drive goal-directed learning and the formation of habit. The differentiation of cortical input to the DMS and DLS in this model is supported by observed cortical projections to the striatum (Hunnicutt et al., 2016; Pan et al., 2010; Peters et al., 2021). Neural activity in the PFC is conceptualized as outcome representations, and the selection of a particular outcome is indicated by the firing rate of corresponding PFC neural populations. This outcome selection is informed by cortico-striatal weights in the DMS, which have been determined by reward prediction error (RPE) over successive trials (Fig. 1B). Neural activity in the PMC corresponds to action representations, and the selection of a particular action is indicated by the firing rate of the corresponding PMC neural populations. This computation integrates outcome representations from the PFC as well as the results of accumulated stimulus-response association in DLS (Fig 1B). The selection of an action determines whether an agent receives rewarding (or punishing) feedback.

This network organization provides the structure for an outcome-to-action map. Since cortico-striatal weights in the DMS track reward prediction errors, the medial partition pairs the neural representations of outcomes with the reward continency of an action. In this way, the PFC-DMS drives behavior by computing desirable outcomes. On the other hand, the lateral partition (PMC-DLS cortico-striatal loops) retrieves what action is most likely to have been paired with the current stimulus. Feedforward coupling from PFC to PMC provides the association of outcome representations and action representations. This connectivity realizes an outcome-to-action map that learns to adapt to new reward feedback rules and receive reward in a consistent and predictable manner.

173 ***Organization of the basal ganglia.***

174 The functional architecture of the BG in this model is derived from our previous models (Kim et
175 al., 2017; Mulcahy et al., 2020), which are in turn informed by Frank (2005). The BG is organized into two
176 subsequent pathways that descend from the striatum: the direct pathway (striatum-SNr/GPi) and the
177 indirect pathway (striatum-GPe-STN-SNr/GPi) (Fig. 1A) (Gerfen and Surmeier, 2011; Gerfen et al., 1990).
178 The direct and indirect pathways converge on the SNr and the GPi, which are the output node of the BG
179 (Alexander and Crutcher, 1990; Hikosaka, 2007; Percheron and Filion, 1991); here, these two regions are
180 treated as the same populations. D1 MSNs directly inhibit the SNr/GPi. D2 MSNs inhibit the GPe which
181 subsequently inhibits the SNr/GPi; in other words, D2 MSN activation leads to disinhibition of the
182 SNr/GPi. Preferential activation of the direct pathway promotes the selection of an action. When D1
183 MSNs are active, they suppress SNr/GPi neurons that would otherwise inhibit the representation of that
184 action in the cortex. On the other hand, preferential activation of the indirect pathway biases against
185 the selection of that action. D2 MSN activation releases SNr/GPi from inhibition, and the corresponding
186 cortical targets are suppressed. Learning in the BG is contingent on the release of dopamine (DA) from
187 the SNc (Schultz, 2007); this reward-based signal induces long-term potentiation in D1 MSNs and long-
188 term depression in D2 MSNs (Calabresi et al., 2007; Shen et al., 2008). By altering the ability of D1 and
189 D2 MSNs to utilize cortical input, DA release supports reward-based learning in this system.

190

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

192 The computations that mediate the selection of the best of the two actions occurred
193 concurrently within the medial and lateral partitions, and we distinguished the results of these two
194 separate processes based on learning modality. Since learning in the medial partition was based on
195 reward prediction error (Lerner et al., 2015) (Table 1), the PFC-DMS weights were updated after each
196 trial to accumulate information about reward contingency (Fig. 1B). As such, we characterized learning

197 in the medial partition as goal-directed learning, and we designated the choice made within the medial
198 partition to be outcome selection. On the other hand, learning in the lateral partition was based on
199 salience of the stimuli to recently received reward. This signal contained information about recent
200 reward history, and the update of PMC-DLS synaptic weights accumulated information about the
201 likelihood that a given stimulus would result in a reward (Fig. 1B). Learning in the lateral partition
202 associated stimuli to rewarding responses. Since the output of the lateral partition defined behavior, we
203 designated the choice made in the lateral partition to be action selection.

204 Outcome selection was independent of action selection, and while these processes typically
205 agreed they would occasionally disagree (Fig. 1C,D). The firing rate of the PFC population determined
206 outcome selection, and each PFC population excited the corresponding PMC population. In this way,
207 outcome selection influenced action selection in the lateral partition. If the PFC population
208 corresponding to outcome #1 was selected, the high firing rate of that PFC population made it more
209 likely that action #1 was selected in the lateral partition, and in some cases both partitions agreed:
210 outcome #1 and action #1 were both selected (Fig. 1C). However, action selection in the lateral partition
211 was also subject to its own learned stimulus-response associations, and the internal dynamics of the
212 lateral partition competed with input from the medial partition to determine which action is performed.
213 As a result, there were cases where the two partitions disagreed: PFC population #2 had a high firing
214 rate indicating the selection of outcome #2, but in the lateral partition, PMC #1 had a high firing rate
215 indicating the selection of action #1 (Fig. 1D).

216

217 ***Reward engages both goal-directed learning and the formation of a stimulus-response association.***

218 Learning in this model was accomplished by changes in the cortico-striatal synaptic weights.
219 Generally, the enhancement of D1 synaptic weights promoted the selection of the corresponding
220 behavioral choice, and the enhancement of D2 synaptic weights made the selection of the

221 corresponding behavioral choice less likely. At the beginning of the initial learning session, D1 and D2
222 weights were not biased to select one of the two options (Fig. 2A; Fig. 2B1 marker ∗1). Over many trials,
223 the reward from the selected action was used to update these weights such that one action was more or
224 less likely to be selected. As discussed below, the medial and lateral partitions utilized reward in
225 different ways.

226 In the medial partition, the reward prediction error (RPE) was used to update cortico-striatal
227 synaptic weights. This quantity could be either positive or negative and change sign from one trial to the
228 next. For example, when there was a low expected reward and the rewarded action (in this case action
229 #1) was chosen, the RPE became positive (Fig. 2A; Fig. 2B2 marker ∗2). If the PFC population
230 representing outcome #1 was active, the PFC-DMS D1 weight for outcome #1 was enhanced and the
231 PFC-DMS D2 weight for outcome #1 was suppressed (Fig. 2A; Fig. 2B3 marker ∗3). Since PFC population
232 #2 was not active at this step, PFC-DMS weights for outcome #2 were not updated based on RPE; these
233 weights slowly decayed to their baseline value.

234 On the other hand, if the unrewarded action (action #2) was chosen, the RPE became negative
235 (Fig. 2A; Fig. 2B4 marker ∗4). If the outcome #1 neurons were also activated in this case, the PFC-DMS
236 D1 weight for outcome #1 was suppressed and the PFC-DMS D2 weight for outcome #1 was enhanced
237 (Fig. 2A; Fig. 2B4-5 marker ∗5). However, if the unrewarded action (action #2) is selected and the PFC
238 selected outcome #2, the PFC-DMS weights were changed to de-emphasize the selection of outcome #2:
239 the D1 weight was decreased, and the D2 weight was increased (Fig. 2A; Fig. 2B4-5 marker ∗6). Over the
240 course of training, the PFC-DMS weights were updated such that the PFC inputs to the DMS channel for
241 outcome #1 efficiently activated the direct pathway, which emphasized the selection of outcome #1,
242 and the PFC inputs to the DMS channel for outcome #2 efficiently activated the indirect pathway, which
243 tended to prevent the selection of outcome #2. Thus, the cortico-striatal weights in the DMS embodied
244 a policy to promote outcome #1 and avoid outcome #2.

Recent data indicates that dopamine release in the DLS encodes behavioral salience rather than reward prediction error (Lerner et al., 2015). Here, we interpret salience to depend on past behavioral engagement and utilize expected reward to compute cortico-striatal weight updates in the lateral partition. 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 weight increased and PMC-DLS D2 weight decreased. In this way, the weights for action #1 ultimately became biased to promote the selection of 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 selected rather than contingency upon outcome.

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

Reward reversal simulations 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

269 selected, the reward prediction error became negative. This selection suppressed the PFC-DMS D1
270 weight for outcome #1, enhanced the PFC-DMS D2 weight for outcome #1, and led to a decrease in the
271 expected reward (Fig. 3A; Fig. 3B1 marker ∗1). However, the PFC-DMS synaptic weights for outcome #2
272 required the selection of action #2 due to its dependence on the reward to be actively updated. Since
273 the PMC-DLS weights were initially biased to select action #1, there was a low likelihood of selecting
274 action #2 for a number of trials.

275 During early reversal trials, synaptic weights in the DMS became biased to favor the indirect
276 pathway for both outcome #1 and outcome #2 (Fig. 3A; Fig. 3B2 marker ∗2), and there was not a strong
277 likelihood for the DMS to emphasize one outcome over the other. The medial partition does not receive
278 input from the lateral partition, so it is free to produce an outcome selection based on its own internal
279 dynamics. Since synaptic weights in the DMS were biased to promote both indirect pathways, the PFC
280 was equally likely to select either outcome #1 or outcome #2 for several trials. In other words, the PFC
281 performed exploration of outcomes. Once the rewarded action #2 was selected, the corresponding PFC-
282 DMS D1 synaptic weight was enhanced; the corresponding PFC-DMS D2 synaptic weight was
283 suppressed; and the subsequent selection rate for outcome #2 increased in the medial partition (Fig. 3A;
284 Fig. 3B3 marker ∗3). The combination of occasional excitatory input to the PMC resulting from the
285 selection of outcome #2 in the PFC and the gradual decay of PMC-DLS synaptic weights that favored
286 action #1 lead to a transition in PMC action selection from the persistent selection of action #1 to
287 exploration of action selection.

288 The decay of PMC-DLS synaptic weights towards the baseline is due to a decrease of the
289 expected reward in early trials. The channel for action #1 in the DLS was less likely to promote action #1,
290 and the decay of these weights contributed to exploration of actions in the lateral partition (Fig. 3A
291 marker ∗4). Concomitantly, goal-directed selection of outcome #2 in the PFC promoted the selection of
292 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. Over several trials, the decay of weights associated with action #1 allowed the emergence of exploratory action selection such that the alternative action #2 could be sampled and then subsequently exploited.

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 value for action #1 to 20% of its 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; Papageorgiou et al., 2016; Tobler et al., 2005).

In this session, the expected reward began close to 1, however, the selection of action #1 provided decreased reward value. The expected reward was large compared to the reward received, 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 outcome #1 were suppressed and PFC-DMS D2 weights for outcome #1 were enhanced in early trials: the channel for outcome #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 outcome #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 value, the expected reward converged to a new value; 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

317 regime, the exploratory phase in outcome selection was abolished. The RPE became positive again when
318 the rewarded action was selected, and the PFC-DMS weights for outcome #1 were restored to select for
319 outcome #1 (Fig. 4A; Fig. 4B3 marker ∗3).

320 The change in magnitude of the reward altered the steady state value of PMC-DLS weights.
321 Since the reward was diminished, the expected reward was also relatively small. The expected reward
322 modulates the magnitude of change in PMC-DLS synaptic weights, so the amount by which action
323 selection changed PMC-DLS weights was relatively small. This active update of synaptic weights was
324 required to prevent decay of weights to their baseline values, and over many trials, the PMC-DLS D1 and
325 D2 weights converged to new values that were closer to baseline synaptic weights. While the stimulus-
326 response association for action #1 was diminished, it was not lost entirely. Following reward
327 devaluation, the PMC-DLS D1 and D2 weights decreased towards baseline but were still configured to
328 promote the selection of action #1. The agent persistently select action #1 with reduced reliability.
329 These results also suggest that the formation of habit is sensitive to the magnitude of the reinforcing
330 stimulus.

331

332 ***Punishment engages goal-directed learning.***

333 The punished outcome session began with cortico-striatal weights inherited from the end of the
334 initial learning session; this model instance was already biased to select action #1. In this session, the
335 magnitude of the reward for action #1 was 50% that in the initial learning session, but it had the
336 opposite valence. We define this negative reward to be punishment, which is based on experiments
337 showing that pauses in DA firing are the dominant response to aversive stimuli (Ilango et al., 2012;
338 Lerner et al., 2015).

339 In early trials of the punished outcome session, the PMC continued to select the previously
340 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 received 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. During these trials, the PFC-DMS D1 and D2 weights for outcome #1 changed valence such that the DMS channel for outcome #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, we altered the feedback used by the DLS to update cortico-striatal synaptic weights. Substantia nigra dopaminergic neurons that project to the DLS respond to both rewarding and aversive stimuli with a positive dopamine transient (Lerner et al., 2015; Matsumoto and Hikosaka, 2009). During punishment sessions, the lateral partition utilized the *rectified* reward expectation to perform weight updates. The rectified expectation integrates the absolute value of the reward. This signal is intended to reflect the behavioral salience of the selection option. In all previous sessions, the reward was positive and the rectified expectation was identical to expected reward. However, in the punished outcome sessions, 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 and deemphasized 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, and action selection was driven by the PFC which sought to minimize the delivery of punishment. This operation minimized the rectified

365 expectation (kept it close to zero). Since the synaptic plasticity update rule for PMC-DLS depends on
366 rectified expectation, which had a small value, the change in PMC-DLS synaptic weights was dominated
367 by the decay of synaptic weights after each trial. Therefore, learning to avoid a punished outcome does
368 not develop the stimulus-response association of choosing an alternative action.
369
370 ***Misalignment of PFC outcome selection with channels in DMS and PMC decreases action selection***
371 ***performance.***
372 Work from our group shows that outcome representations are reduced in animals that drink
373 alcohol compulsively, and also demonstrate impaired behavioral flexibility (De Falco et al., 2021;
374 Linsenbardt et al., 2019; Timme et al., 2021). More specifically, reductions in mutual information about
375 two competing options was observed in neural recordings obtained from the mPFC of rats that drink
376 alcohol compulsively, indicating that the representation of these options was not clearly disambiguated.
377 To simulate this observation, we repeated reward reversal and punished outcome simulations
378 incorporating reduced outcome coding in the PFC (Fig. 6). In these simulations, each PFC population
379 routed 90% of its projections to the appropriate channel and 10% of its projections to the wrong
380 channel (striatal neurons corresponding to the other outcome) (Fig. 6A). The inhibitory input from the
381 GPi populations to the PFC was similarly mis-matched to retain balance between excitation and
382 inhibition in these networks.
383 These changes in connectivity in the model are intended to reflect a change in the fidelity of
384 prefrontal-cortical representations. The output of each PFC channel to the DMS and the PMC, and the
385 feedback to the PFC from the GPi is contaminated by information from the opposing channel. In this way
386 this impairment lead to a PFC outcome signal that is slightly contaminated by the alternative outcome.
387 We interpret this implementation of impaired executive control as a disruption of the outcome-to-
388 action mapping. The mechanisms for goal-directed learning in the DMS operate on representations that

do not cleanly map to representations in the PFC, and the PFC projections into the PMC do not cleanly promote either action #1 or action #2. In simulations of both reward reversal and punishment, impaired 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 impaired outcome coding in the PFC, 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 was 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 simulations of reward reversal and reward reversal with the impaired outcome coding in the PFC ($\mathrm{PFC}^{\dagger}$) and between punished outcome and punished outcome $\mathrm{PFC}^{\dagger}$ sessions (Fig. 7B). From the reward reversal to the reward reversal $\mathrm{PFC}^{\dagger}$ sessions, the median agent change point significantly increased from trial 15 to 21 in outcome in the PFC (Mann-Whitney U = 2441.5, n_1=n_2=100, p=3.90E-10) and from trial 47.5 to 85 in the PMC (Mann-Whitney U = 134.5, n_1=n_2=100, p=1.35E-32). From punished outcome to the punished outcome $\mathrm{PFC}^{\dagger}$ sessions, the median agent change point significantly increased from trial 16 to 22 in the PFC (Mann-Whitney U = 3705.5, n_1=n_2=100, p=1.56E-3) and from trial 81 to 162.5 in the PMC (Mann-Whitney U = 87.5, n_1=n_2=100, p=3.41E-33). 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 2). 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. In the devaluation session, there was no significant difference between PFC and PMC performance (Table 2). In punished outcome sessions, the PMC was less likely to choose the unpunished action than the PFC was to choose the corresponding outcome (Table 2). We also observed a significant difference in steady state performance of action selection between initial learning and devaluation sessions (Mann-Whitney U = 10000, $n_1=n_2=100$,p=2.2E-16)

Session Name	Mann-Whitney U statistic	p-value
Initial learning	1958.5	1.1E-13
Devaluation	4070	0.023
Reward reversal	8333.5	3.8E-16
Punished outcome	1737.5	1.6E-15
Reward Reversal $\mathrm{PFC}^{\dagger}$	8352.5	2.6E-16
Punished outcome $\mathrm{PFC}^{\dagger}$	907.0	1.5E-23

Table 2. 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, the selection of the previously rewarded action perseverated even after cortico-striatal weights in the DMS were configured in a manner to support exploration. Specifically, DLS-driven action selection countered goal-directed behavior despite reward reversal, reward devaluation, and the punishment of an outcome, and in these scenarios, we characterized behavior as habitual.

In this work, we utilized the anatomical and physiological segregation of dopamine signals in the striatum to demonstrate mechanisms that drive separate processes for goal-directed and habitual learning. Our work follows recent investigations that approach similar problems exploring how behavior is influenced by cortico-basal ganglia interactions. Miller et al. (2019) has recently challenged previous models for habit in reinforcement learning. Instead, the authors propose a value-free habit mechanism to describe stimulus-response associations based on Thorndike's second Law of Exercise—behaviors that have been repeated often in the past are more likely to be performed in the future. 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. 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, which will be critical to address in future iterations of our model. Bogacz (Bogacz, 2020) has developed a mathematical framework to describe goal-directed learning and habit as a process based on Bayesian inference. Their algorithm is broken down into basic units that are compared to the anatomy and physiology of the basal ganglia including separate components for the DMS and DLS. This framework differs from our own

model by including circuitry that estimates and controls for uncertainty to weight the output of the goal-directed and habit systems (Bogacz, 2020). Baladron and Hamker (Baladron and Hamker, 2020) have implemented simulations of the neural circuitry in the DMS and DLS respectively as goal-directed and stimulus-response controllers. Their modeling framework differs from our own by storing the synaptic weights responsible for habitual behavior in the infralimbic cortex, and these weights are slowly updated based on a dopamine-independent Hebbian plasticity learning rule (Baladron and Hamker, 2020). We extend this body of work by demonstrating how a lack of fidelity in neural representations at the level of the PFC could facilitate the dominance of habitual responding. We accomplish this result by implementing an impairment of outcome representations in the PFC, which is dynamically involved in learning in our model of cortico-striatal interactions. Here, we implemented 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 in the face of drug seeking (Corbit et al., 2012; Murray et al., 2012). 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 representations of outcomes in the PFC fail to map cleanly to actions, then goal-directed learning in the DMS will lose efficacy, and agents will take longer to overcome previously learned stimulus-response associations. We implemented this impairment of executive control by slightly contaminating outcome representations with their alternatives in devaluation, reward reversal, and punished outcome tasks. Agents with impaired PFC took longer to learn new reward feedback rules and exhibited worse steady state performance.

479

480 ***Reward devaluation and punishment reveal mechanisms that support habitual behavior***

481 Experimental protocols may induce reward devaluation through manipulations that remove
482 drive, for example by selective satiety, or induce an aversive association with the reward, for example
483 using lithium chloride (Rossi and Yin, 2012). In our model design, we distinguish between reward
484 devaluation tasks—which we accomplish by directly reducing the quantitative value of the reward for
485 action #1, and punished outcome tasks—in which the valence of the feedback associated with action #1
486 changes from positive to negative. Both of these protocols provide the opportunity to evaluate the role
487 of learning in the DMS and the DLS to overcome an established stimulus-response association. Through
488 direct inspection of the models neural activity and cortico-striatal synaptic weights, we have shown,
489 during the trials immediately following a change in outcome contingency, how stimulus-response
490 associations drive behavior in spite of rapid goal-directed learning and updated reward-outcome
491 contingencies in the PFC neural representations. These dynamics are an illustration of a mechanism by
492 which neuronal circuitry that is critical for habitual responding resists or overpowers the action of
493 circuitry that is critical for goal-directed behavior, and we classify the inflexible behavior driven by these
494 dynamics as habit.

495 Punishment simulations were motivated in part by experiments in rodents that included a foot
496 shock stimulus (Lerner et al., 2015). The reward-based feedback communicated via DA to the DLS—the
497 rectified expectation—integrates the absolute value of the trial reward and is distinct from the expected
498 reward. In this way, the dopaminergic signal to the DLS is positive for either a rewarding or an aversive
499 stimulus (as in Lerner et al., 2015). In goal-directed behavior, the aversive choice is avoided; in this case
500 the agent seeks to minimize the negative reward prediction error and, hence, the quantity of negative
501 expected reward. Since the rectified expectation is minimized by the learning process, no new stimulus-
502 response association is able to develop in punished outcome tasks. The results of learning in punished

503 outcome tasks differ from the results of reward reversal. In both cases, the stimulus-response
504 association for a previously rewarded action must be overcome. In the reward reversal session, goal-
505 directed behavior works to maximize the reward delivered for the newly rewarded action. Hence, the
506 expected reward is maintained at a large value over the course of training, and a new stimulus-response
507 association is formed. This critical difference between reward reversal and punished outcome tasks—
508 whether there is a new stimulus-response association or not—is visualized in the comparison of steady
509 state performance (Fig. 7B). In reward reversal simulations, the performance of outcome selection in the
510 PFC is not as great as action selection in the PMC (discussed below). Thus, steady state behavior is
511 driven by the new stimulus-response association formed in the DLS. However, in punished outcome
512 tasks, steady state behavior is maintained by goal-directed learning (the medial cortico-striatal partition)
513 in the absence of a new stimulus-response association in the DLS. The performance of outcome
514 selection in the PFC leads performance of action selection in the PMC (Fig. 7B). These results lead to the
515 hypothesis that the control of behavior cannot be acquired by DLS, thereby coming under habitual
516 control, in tasks where punishment is used as a reinforcer, and therefore habits will not form. These
517 results can also be examined in the context our reward devaluation sessions. Here, the reward
518 magnitude is a fraction of that during initial learning. Despite this change, the small reinforcer is still
519 sufficient to engage learning in the DLS. Even though responding is less reliable in reward devaluation
520 sessions than initial learning sessions, responding resists dramatic changes D1 and D2 weights of the
521 DMS during early trials that would otherwise indicate reversal and selection of the alternate outcome
522 (Fig. 4AB).
523
524 ***PFC coding failure exacerbates inflexible behavior***
525 Our implementation of impaired executive control is intended to represent a failure of outcome
526 representations in the PFC to appropriately map to action representations in the PMC. The prefrontal-

527 cortical input to channels for outcome #1 and outcome #2 in the DMS and the PMC populations for
528 action #1 and action #2 were slightly mixed in $\mathrm{PFC}^{\dagger}$ sessions. Specifically, the input to the DMS channel
529 for outcome #1 and the input to the PMC population for action #1 was a weighted some of both PFC
530 populations: 90% PFC for outcome #1 + 10% PFC for outcome #2 (Fig. 6A). The firing rates of these
531 populations at the end of the simulation were used to update cortico-striatal synaptic weights in the
532 DMS between trials of the session (see Methods). During selection of an outcome in the PFC, the PFC
533 population corresponding to the selected outcome had a high firing rate, and the PFC population
534 corresponding to the unselected outcome had a low firing rate (Fig. 1 C,D). Due to the manipulation, the
535 firing rate used to update synaptic weights for the selected outcome is slightly lesser compared to
536 control simulations, resulting in a slightly lesser change in synaptic weights. On the other hand, the firing
537 rate used to update synaptic weights for the unselected outcome is slightly greater compared to control
538 simulations, resulting in an inappropriately large change in the corresponding synaptic weights.
539 Similarly, the excitatory input from PFC to PMC was less distinct between channels: the PFC was slightly
540 less able to promote the selection of the appropriate action in the PMC in $\mathrm{PFC}^{\dagger}$sessions compared to
541 control simulations.
542 Our simulations of PFC impairment are intended to illustrate how impaired executive control in
543 the PFC could promote the emergence of inflexible behavior. The PFC is critical for adaptive behavioral
544 control, and impairment of the PFC disrupts cognitive flexibility (Jobson et al., 2021). Our results show a
545 mechanism by which the impairment of the PFC could reduce goal-directed responding and facilitate the
546 emergence of inflexible behavior. After loss of executive control, agents have reduced performance in
547 simulations with altered action-outcome contingencies, and these differences appear in both the
548 outcome representations in the PFC and in action-selection performance as indicated by PMC (Fig. 7B).
549 This deficiency in cognitive flexibility can be related to behavioral performance of animals with PFC

550 dysfunction. Deficits in reversal learning performance have been associated with lesions or alcohol-
551 induced damage to the orbitofrontal cortex (Dias et al., 1997; Izquierdo et al., 2017; Moorman, 2018).
552 How could repeated exposure to alcohol change the brain to induce a failure for neural activity
553 to properly code for action selection? Dopamine dynamics in the PFC are critical for higher functions and
554 executive control (Ott and Nieder, 2019; Puig et al., 2014), and changes to dopamine modulation in the
555 PFC are concurrent with alcohol-induced deficits in executive function (Trantham-Davidson and
556 Chandler, 2015). Patients that suffer from a history of alcohol abuse may have deficient expression of D2
557 dopamine receptors in the PFC (Volkow et al., 1996, 2002). In rats, chronic exposure to alcohol disrupts
558 D2 receptor function in the PFC (Trantham-Davidson et al., 2014). It is thought that D1 receptors in the
559 PFC act to stabilize network activity and D2 receptors promote flexibility (Durstewitz et al., 2000; Jenni
560 et al., 2017). Disruptions in the PFC dopamine system that favor D1 receptors may render an individual
561 unable to quickly adapt in the face of changing behavioral context.
562
563 ***The role and organization of direct and indirect pathways***
564 In reversal learning and punishment follow-up sessions, we observe two prominent phases in
565 the evolution of cortico-striatal weights in the DMS. First, the BG channel for the previously rewarded
566 action becomes essentially inactivated. The cortico-striatal weights for D1 MSNs decreased in value
567 below baseline, and the weights for D2 MSNs increased above baseline. During this phase, the model
568 performance changed to become less likely to select the previously rewarded action, and as a result, it
569 explored alternative actions—here there was only one alternative action. Second, this exploration phase
570 was terminated by the potentiation of the cortico-striatal weight for D1 and depression of D2 MSNs that
571 correspond to the newly rewarded action. A prediction that emerges from these results is that D2 MSNs
572 are critical for the transition to exploratory behavior. This prediction is supported by recent results

573 showing that after initial learning the behavioral flexibility is reduced in a task that requires learning new
574 action-outcome associations by optogenetic inhibition of indirect pathway MSNs (Peak et al., 2020).
575 The observation that both direct and indirect pathway MSNs are activated at the onset of
576 movement supports the hypothesis that direct and indirect pathways are concurrently involved in
577 selection decisions (Cui et al., 2013; Tecuapetla et al., 2016). In the "complementary" model of the
578 striatum, clusters of direct pathway MSNs are selectively activated while indirect pathway MSNs broadly
579 inhibit competing clusters (Bariselli et al., 2019). In the present model, each action is represented by a
580 cluster of direct pathway MSNs within the striatum. However, the present model differs from the
581 "complementary" model in that each cluster of direct pathway MSNs acts in opposition to a selective
582 cluster of indirect pathway MSNs. The computation performed in the SNr/GPi output node of our model
583 is based on competition between the direct and indirect pathways associated with each action; in other
584 words, the BG channel for each action emits a GO or a NOGO signal. In this way, our computational
585 model is most like the "competitive" model described in (Bariselli et al., 2019) where direct and indirect
586 pathway MSNs compete within an ensemble to promote or inhibit an action. Our model uses multiple
587 channels to represent distinct action options, and competition between channels—and hence between
588 actions—is implemented via mutual inhibition within each cortical compartment.
589
590 ***Striatal Dopamine Distribution***
591 The spatio-temporal distribution of dopamine release within the striatum is critical for selective
592 reward credit assignment (Hamid et al., 2021). However, dopamine release supports different learning
593 modalities in different compartments of the striatum (Lerner et al., 2015; Matsumoto and Hikosaka,
594 2009). Moreover, dopamine release occurs in wave-like patterns that propagate in the medial-lateral
595 axis of the dorsal striatum, and the direction of wave propagation depends on behavioral context
596 (Hamid et al., 2021). Specifically, tasks in which reward depended on animal performance promoted

597 dopamine release beginning in the dorsomedial striatum, and tasks in which reward was independent of
598 animal behavior promoted dopamine release beginning in the dorsolateral striatum. The authors (Hamid
599 et al., 2021) performed computer simulations to illustrate the hypothesis that spatiotemporal gradients
600 in dopamine release could emphasize reward credit assignment to specialized striatal subregions and
601 recruit behavioral strategies appropriate for the task at hand. Temporal coincidence in the activity of
602 multiple signals may be involved in functional segregation of the striatal subregions (Aosaki et al., 1994;
603 Morris et al., 2004). A joint consideration of both spatio-temporal (Hamid et al., 2021) and modally
604 (Lerner et al., 2015; Matsumoto and Hikosaka, 2009) mixed dopamine signals could further differentiate
605 the function of different compartments of the striatum.
606
607 ***Predictions and Limitations***
608 A key feature of learning is the transition from exploration to persistently performing a
609 rewarding behavior. Here, we show via simulations how different learning modalities across the DMS
610 and DLS could interact to contribute to this transition. The balance between exploration of unknown
611 options and exploitation of known rewards is a ubiquitous challenge necessary for the survival of
612 animals (Wilson et al., 2021). Even after learning new reward-feedback rules, animals sample alternate
613 actions, and this alternative sampling is representative of continual exploration (Ebitz et al., 2019;
614 Pisupati et al., 2021). The neurophysiological basis for exploration and strategic control is an active topic
615 of research with focus in the cortex (Cervera et al., 2020; Kaanders et al., 2021).
616 Our results here suggest a mechanism for the involvement of the DMS in the support of
617 alternative sampling after a reward contingency rule has been learned. Once behavior becomes driven
618 primarily by stimulus-response association—in other words by the DLS in the lateral partition—cortico-
619 striatal synaptic weights in the DMS are not required to maintain agent behavioral performance. What
620 role, then does outcome selection in the medial partition play after extensive training? At steady state,

621 we observed a higher error rate in the outcome representations of the PFC than in the action selection
622 as indicated by PMC (Fig. 7C). Mistakes or lapses in the medial partition increase the likelihood of
623 alternative sampling in action selection. We formulate this observation as a key prediction of our
624 computational model: neuronal activity in the PFC should occasionally and increasingly correspond to
625 alternative unrewarded actions as an animal learns to persistently select a rewarded action. In this way,
626 the PFC and DMS could monitor alternative reward contingencies and contribute to tonic exploration.

627 Drug use or risk of drug use is associated with impairments in cognitive flexibility (Barker et al.,
628 2017; Ortega et al., 2013; Perry and Carroll, 2008; Shnitko et al., 2019; STALNAKER et al., 2007).
629 Following chronic exposure to drugs, animals are more likely to make perseverative errors immediately
630 following a change in reward rule (Ebitz et al., 2019; Kroener et al., 2012). We relate these results to our
631 reward reversal $\mathrm{PFC}^{\dagger}$ simulations: in simulations with impaired outcome coding in the PFC, agents
632 exhibit a substantial increase in perseverative errors (Fig. 7A). Since model behavior in early trials was
633 dependent on goal-directed learning, the impairment of outcome coding in the PFC compromised the
634 ability of agents to overcome stimulus-response associations for previously rewarded actions.

635 Our model implements the cortex in a limited fashion, and in future work, this aspect of the
636 model could be extended to capture greater anatomical and physiological detail. In our previous models,
637 stimulus-response associations were captured by Hebbian learning in cortico-cortical synaptic weights
638 (Kim et al., 2017; Mulcahy et al., 2020). Plasticity in these projections was omitted in the present study
639 to emphasize the role of DLS in the acquisition of stimulus-response associations and habit. However,
640 cortical mechanisms for learning are critical for cognitive and executive flexibility (Floresco, 2013). The
641 inclusion of cortical learning would permit contingency representations external to the striatum and
642 allow us to investigate the cortical dopamine dynamics of learning tasks in healthy animals and in
643 animals that suffer from the chronic use of addictive substances. Additionally, our present model does
644 not distinguish between different regions of the prefrontal cortex. The prelimbic and infralimbic cortices

645 appear to play opposing roles in learning (Amaya and Smith, 2018; Lipton et al., 2019). Disruption of
646 infralimbic cortex prevents the expression of habit and facilitates goal-directed behavior (Coutureau and
647 Killcross, 2003; Smith et al., 2012), and the interaction of the prelimbic and infralimbic regions is critical
648 for the performance of set shifting tasks (Mukherjee and Caroni, 2018). Expanding this model to include
649 the specific functions of these brain regions may accommodate aspects of behavior where learning
650 resembles an immediate state change (e.g. attentional set shifting) as opposed to gradual changes over
651 trials.
652

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.

$$I_{D1} = g_{CTX} w_{CTX-D1} CTX$$

$$I_{D2} = g_{CTX} w_{CTX-D2} CTX$$

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

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

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

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 impaired outcome coding in the PFC

In $\mathrm{PFC}^{\dagger}$ sessions, we alter the connectivity to and from the PFC in order to simulate scenarios in which the representations of outcomes in the PFC do not cleanly map to representations of actions in

the PMC. 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 3.

Synaptic plasticity.

Dopaminergic input to the DMS codes the reward prediction error, which is the difference between reward (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. The coefficient α determines how new reward is waited against previous expected reward to determine the next value for expected reward. 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 is used to compute the rectified expectation ($RectExp$).

$$SNc_{DLS} = RectExp_j$$

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

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 3. 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 would be delivered, and this reward 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 with different random number generator seeds, 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 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, punished outcome, reward reversal $PFC^{\dagger}$, and punished 35utcome $PFC^{\dagger}$ (see Table 4). 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 for action #1 was set to 0.2. In reward reversal sessions, the reward was set to 0 for action #1 and 1 for action #2. In the punished outcome session, the reward was set to -0.5 for action #1 and 0 for action #2. The reward for the reward reversal $PFC^{\dagger}$ and punished outcome $PFC^{\dagger}$ sessions were identical to the reward reversal and punished outcome 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
Punished outcome	-0.5	0	2000
Reward reversal $\mathrm{PFC}^{\dagger}$	0	1	2000
Punished outcome $\mathrm{PFC}^{\dagger}$	-0.5	0	2000

Table 4. Behavioral sessions were defined by the organization of reward feedback.

Evaluating model performance.

The group performance of agents in reward reversal and punished outcome 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 $1/201$ as 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 punished outcome sessions to reward reversal $\mathrm{PFC}^{\dagger}$ and punished outcome $\mathrm{PFC}^{\dagger}$ 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 adjusted using the Bonferroni correction for four comparisons (0.05/4 or exactly 0.0125).

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 using the Mann-Whitney U test with an alpha level determined by the Bonferroni correction for seven comparisons; the alpha value was 0.05 / 7 (approximately 7.1E-3). 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

Alexander, G.E., and Crutcher, M.D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci *13*, 266–271.

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.

Baladron, J., and Hamker, F.H. (2020). Habit learning in hierarchical cortex-basal ganglia loops. Eur J Neurosci *52*, 4613–4638.

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.

Barker, J.M., Bryant, K.G., Osborne, J.I., and Chandler, L.J. (2017). Age and Sex Interact to Mediate the Effects of Intermittent, High-Dose Ethanol Exposure on Behavioral Flexibility. Front Pharmacol *8*, 450.

Bogacz, R. (2020). Dopamine role in learning and action inference. Elife *9*, e53262.

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.

Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends in Neurosciences *30*, 211–219.

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.

Dias, R., Robbins, T.W., and Roberts, A.C. (1997). Dissociable Forms of Inhibitory Control within Prefrontal Cortex with an Analog of the Wisconsin Card Sort Test: Restriction to Novel Situations and Independence from “On-Line” Processing. J Neurosci *17*, 9285–9297.

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.

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

Gerfen, C.R., Engber, T.M., Mahan, L.C., Susel, Z., Chase, T.N., Monsma, F.J., and Sibley, D.R. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. Science *250*, 1429–1432.

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.

Hikosaka, O. (2007). GABAergic output of the basal ganglia. Prog Brain Res *160*, 209–226.

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.

Ilango, A., Shumake, J., Wetzel, W., Scheich, H., and Ohl, F.W. (2012). The Role of Dopamine in the Context of Aversive Stimuli with Particular Reference to Acoustically Signaled Avoidance Learning. Frontiers in Neuroscience *6*.

Izquierdo, A., Brigman, J.L., Radke, A.K., Rudebeck, P.H., and Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. Neuroscience *345*, 12–26.

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.

Jobson, D.D., Hase, Y., Clarkson, A.N., and Kalaria, R.N. (2021). The role of the medial prefrontal cortex in cognition, ageing and dementia. Brain Commun *3*, fcab125.

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*.

Klaus, A., Martins, G.J., Paixao, V.B., Zhou, P., Paninski, L., and Costa, R.M. (2017). The Spatiotemporal Organization of the Striatum Encodes Action Space. Neuron *95*, 1171-1180.e7.

Kroener, S., Mulholland, P.J., New, N.N., Gass, J.T., Becker, H.C., and Chandler, L.J. (2012). Chronic Alcohol Exposure Alters Behavioral and Synaptic Plasticity of the Rodent Prefrontal Cortex. PLoS One *7*, e37541.

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.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

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.

Moorman, D.E. (2018). The role of the orbitofrontal cortex in alcohol use, abuse, and dependence. Progress in Neuro-Psychopharmacology and Biological Psychiatry *87*, 85–107.
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.
Niv, Y. (2019). Learning task-state representations. Nat Neurosci *22*, 1544–1553.
Ortega, L.A., Tracy, B.A., Gould, T.J., and Parikh, V. (2013). Effects of chronic low- and high-dose nicotine on cognitive flexibility in C57BL/6J mice. Behav Brain Res *238*, 134–145.
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.
Papageorgiou, G.K., Baudonnat, M., Cucca, F., and Walton, M.E. (2016). Mesolimbic Dopamine Encodes Prediction Errors in a State-Dependent Manner. Cell Rep *15*, 221–228.
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.
Peak, J., Chieng, B., Hart, G., and Balleine, B.W. (2020). Striatal direct and indirect pathway neurons differentially control the encoding and updating of goal-directed learning. Elife *9*, e58544.
Percheron, G., and Filion, M. (1991). Parallel processing in the basal ganglia: up to a point. Trends Neurosci *14*, 55–59.
Perry, J.L., and Carroll, M.E. (2008). The role of impulsive behavior in drug abuse. Psychopharmacology (Berl) *200*, 1–26.
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*.

971 Puig, M.V., Rose, J., Schmidt, R., and Freund, N. (2014). Dopamine modulation of learning and
972 memory in the prefrontal cortex: insights from studies in primates, rodents, and birds. Front.
973 Neural Circuits *8*.
974 Redish, A.D. (2016). Vicarious trial and error. Nat Rev Neurosci *17*, 147–159.
975 Rossi, M.A., and Yin, H.H. (2012). Methods for studying habitual behavior in mice. Curr Protoc
976 Neurosci *Chapter 8*, Unit 8.29.
977 Schacht, J.P., Anton, R.F., and Myrick, H. (2013). Functional neuroimaging studies of alcohol
978 cue reactivity: a quantitative meta-analysis and systematic review. Addict Biol *18*, 121–133.
979 Schultz, W. (2007). Behavioral dopamine signals. Trends in Neurosciences *30*, 203–210.
980 Schwabe, L., and Wolf, O.T. (2011). Stress-induced modulation of instrumental behavior: from
981 goal-directed to habitual control of action. Behav Brain Res *219*, 321–328.
982 Shen, W., Flajolet, M., Greengard, P., and Surmeier, D.J. (2008). Dichotomous Dopaminergic
983 Control of Striatal Synaptic Plasticity. Science *321*, 848–851.
984 Shnitko, T.A., Gonzales, S.W., and Grant, K.A. (2019). Low cognitive flexibility as a risk for
985 heavy alcohol drinking in non-human primates. Alcohol *74*, 95–104.
986 Smith, K.S., Virkud, A., Deisseroth, K., and Graybiel, A.M. (2012). Reversible online control of
987 habitual behavior by optogenetic perturbation of medial prefrontal cortex. Proc Natl Acad Sci U
988 S A *109*, 18932–18937.
989 STALNAKER, T.A., ROESCH, M.R., CALU, D.J., BURKE, K.A., SINGH, T., and
990 SCHOENBAUM, G. (2007). Neural Correlates of Inflexible Behavior in the Orbitofrontal–
991 Amygdalar Circuit after Cocaine Exposure. Ann N Y Acad Sci *1121*, 598–609.
992 Tecuapetla, F., Jin, X., Lima, S.Q., and Costa, R.M. (2016). Complementary Contributions of
993 Striatal Projection Pathways to Action Initiation and Execution. Cell *166*, 703–715.
994 Timme, N.M., Ma, B., Linsenbardt, D., Cornwell, E., Galbari, T., and Lapish, C. (2021).
995 Compulsive drinking is associated with neural activity patterns reflecting diminished behavioral
996 control and enhanced seeking representations in dorsal medial prefrontal cortex. BioRxiv
997 2021.03.15.435169.
998 Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive Coding of Reward Value by
999 Dopamine Neurons. Science *307*, 1642–1645.
1000 Trantham-Davidson, H., and Chandler, L.J. (2015). Alcohol-induced alterations in dopamine
1001 modulation of prefrontal activity. Alcohol *49*, 773–779.
1002 Trantham-Davidson, H., Burnett, E.J., Gass, J.T., Lopez, M.F., Mulholland, P.J., Centanni,
1003 S.W., Floresco, S.B., and Chandler, L.J. (2014). Chronic alcohol disrupts dopamine receptor
1004 activity and the cognitive function of the medial prefrontal cortex. J Neurosci *34*, 3706–3718.
1005 Tran-Tu-Yen, D.A.S., Marchand, A.R., Pape, J.-R., Di Scala, G., and Coutureau, E. (2009).
1006 Transient role of the rat prelimbic cortex in goal-directed behaviour. Eur J Neurosci *30*, 464–
1007 471.
1008 Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S., and Schultz, W. (2016). A dynamic code for
1009 economic object valuation in prefrontal cortex neurons. Nat Commun *7*, 12554.
1010 Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M.M., Uchida, N., and Watabe-Uchida, M.
1011 (2020). Distinct temporal difference error signals in dopamine axons in three regions of the
1012 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.

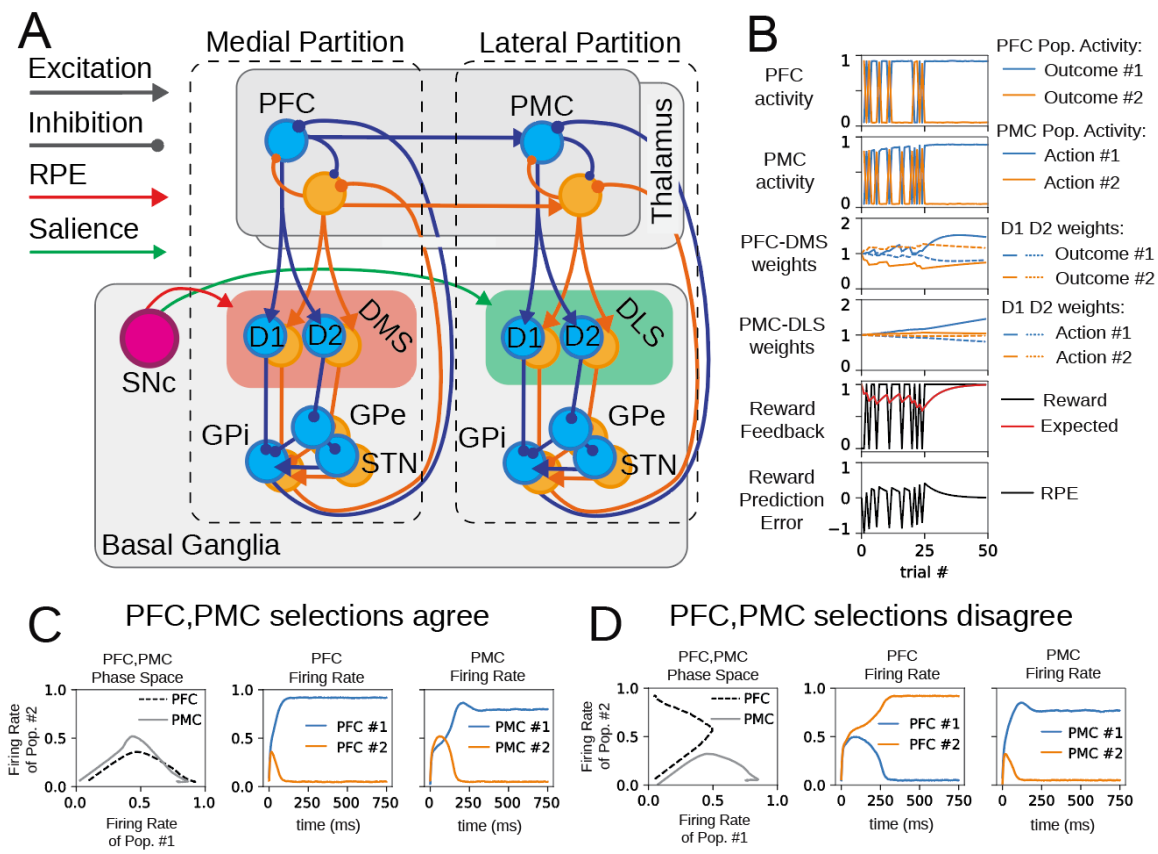

Figure 1. A model of bimodal and concurrent learning in the basal ganglia.
(A) Cortico-basal ganglia-thalamo-cortical loops are distributed between the medial and lateral partitions of the model and are organized into distinct channels that represent individual actions (color-coded as blue and orange). 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).

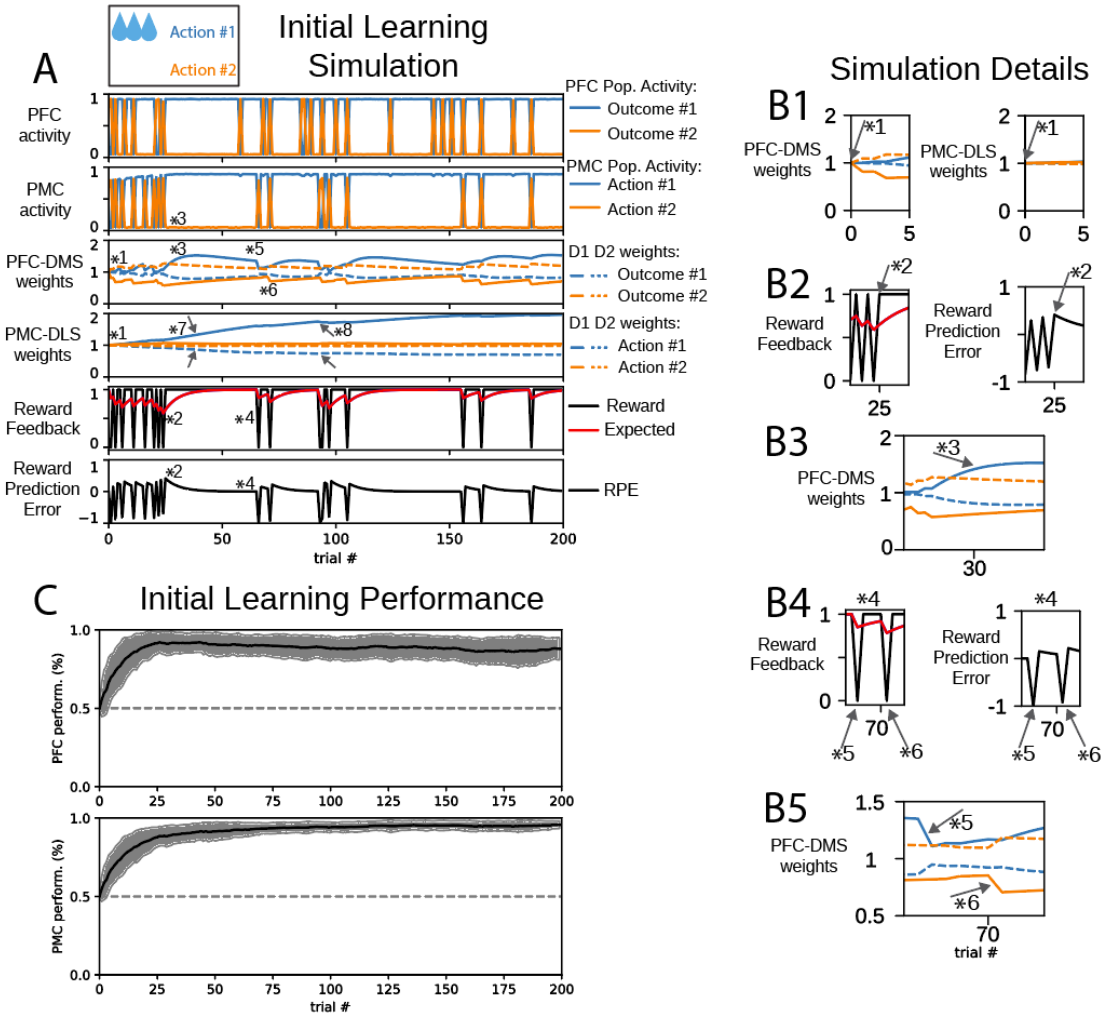

Figure 2.

(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.

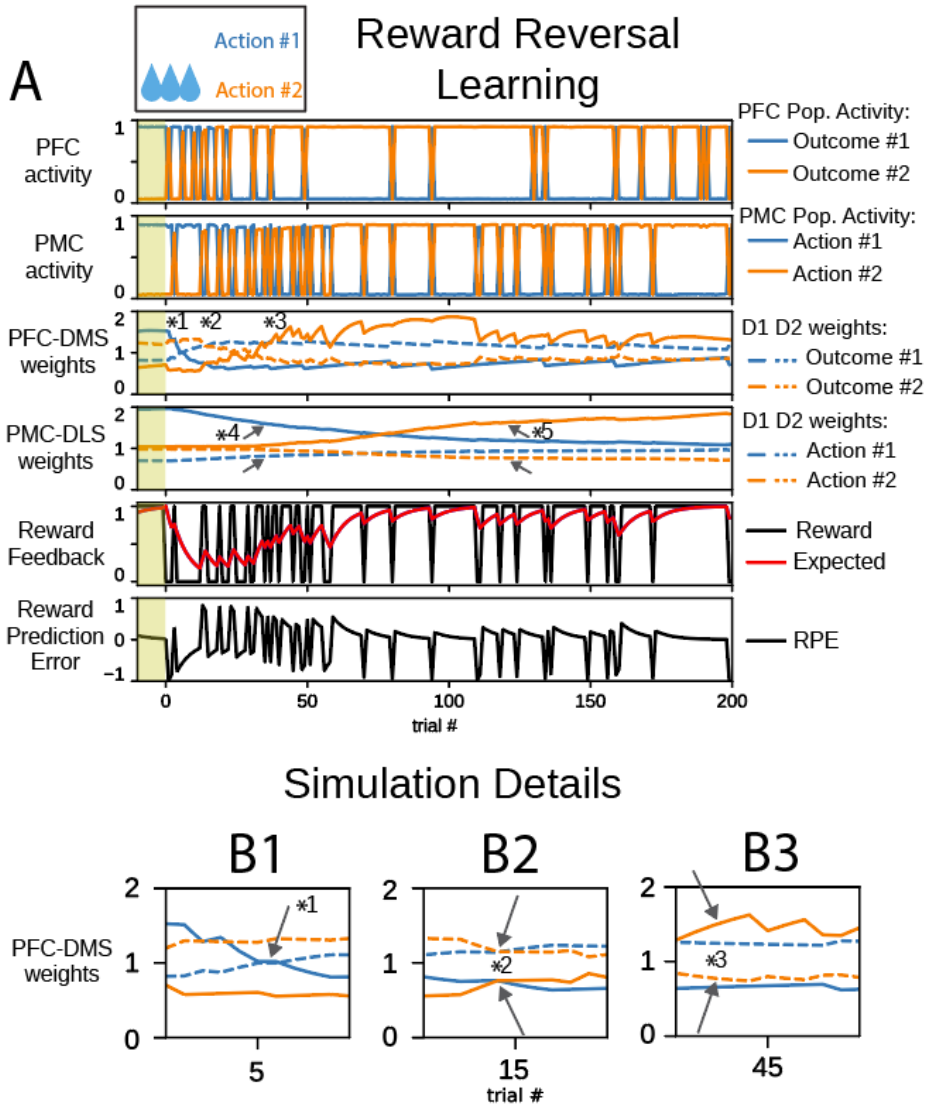

Figure 3. An agent learns to select action #2 in a reward reversal behavioral task.
(A) In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond to action #1 and orange traces correspond to action #2. The yellow 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.

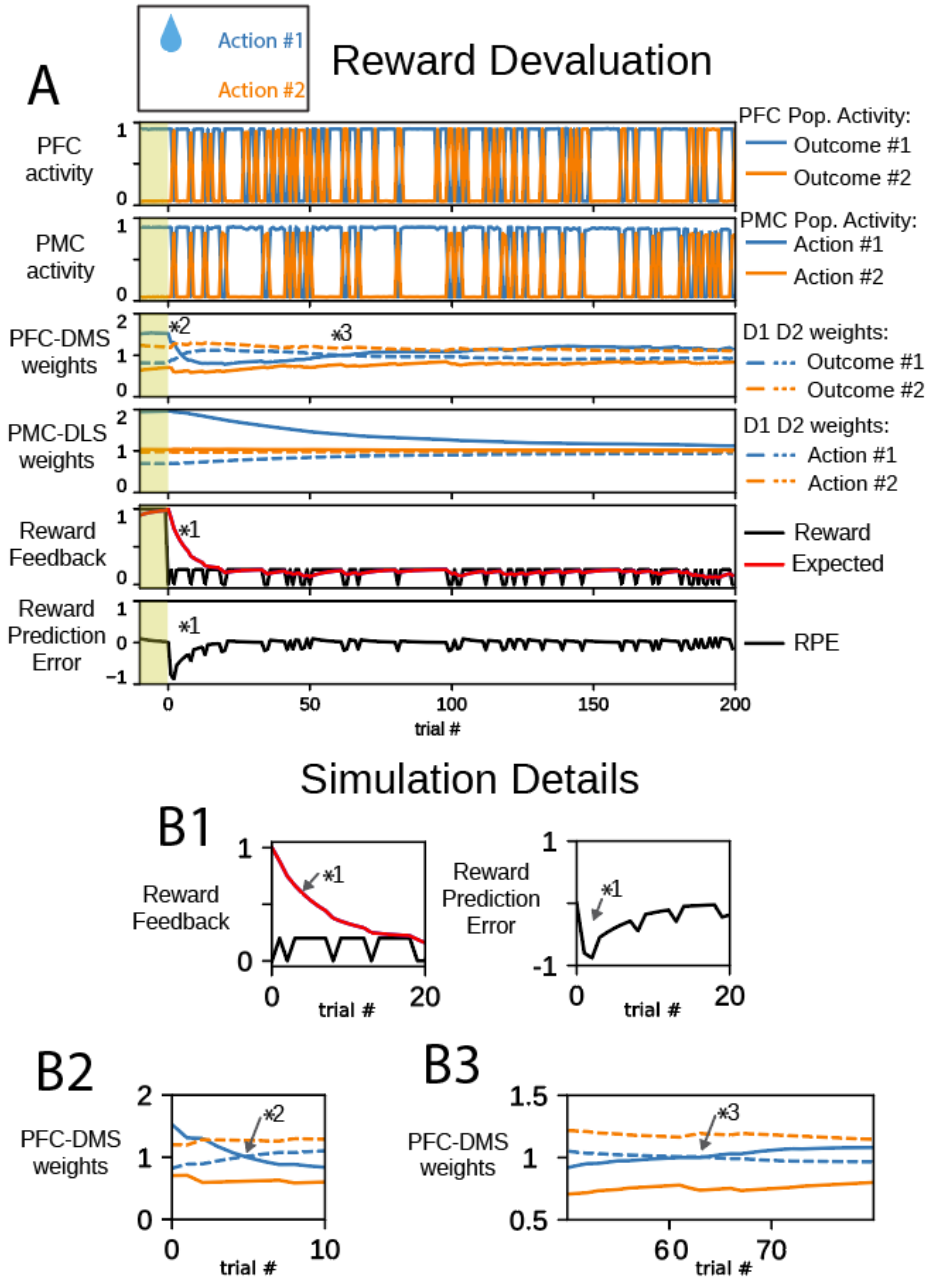

Figure 4. The magnitude of reward feedback is reduced in a reward devaluation behavioral task. The selection of action #1 persists but becomes less likely.
(A) In panels depicting cortico-striatal weights, D1 synaptic weights are solid traces and D2 synaptic weights are dashed traces. Blue traces correspond 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.

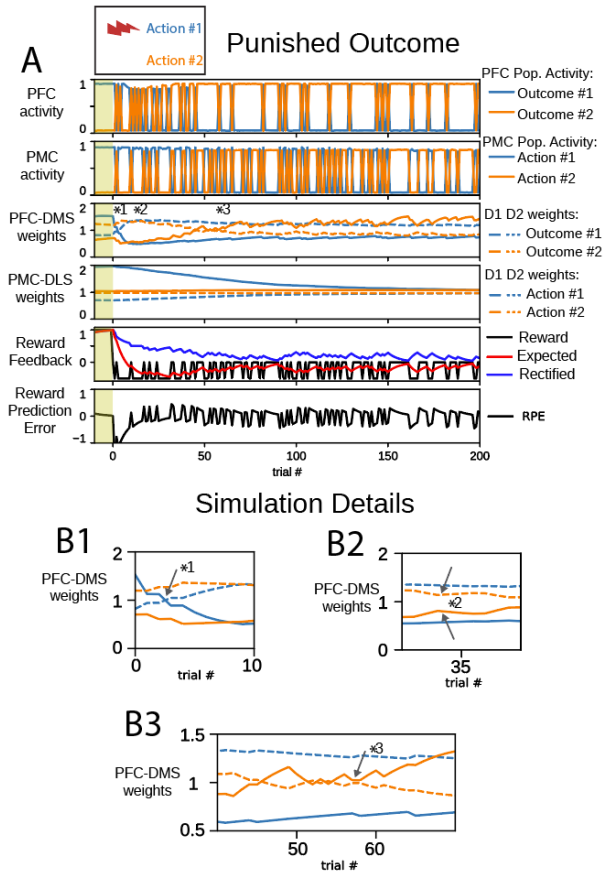

Figure 5. An agent learns to select action #2 in a punished outcome behavioral task.
Cortical activity, cortico-striatal synaptic weights, and reward feedback metrics evolve to overcome habit to select action #1. The yellow shaded trials before trial 0 indicate trials at the end of the initial learning 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.

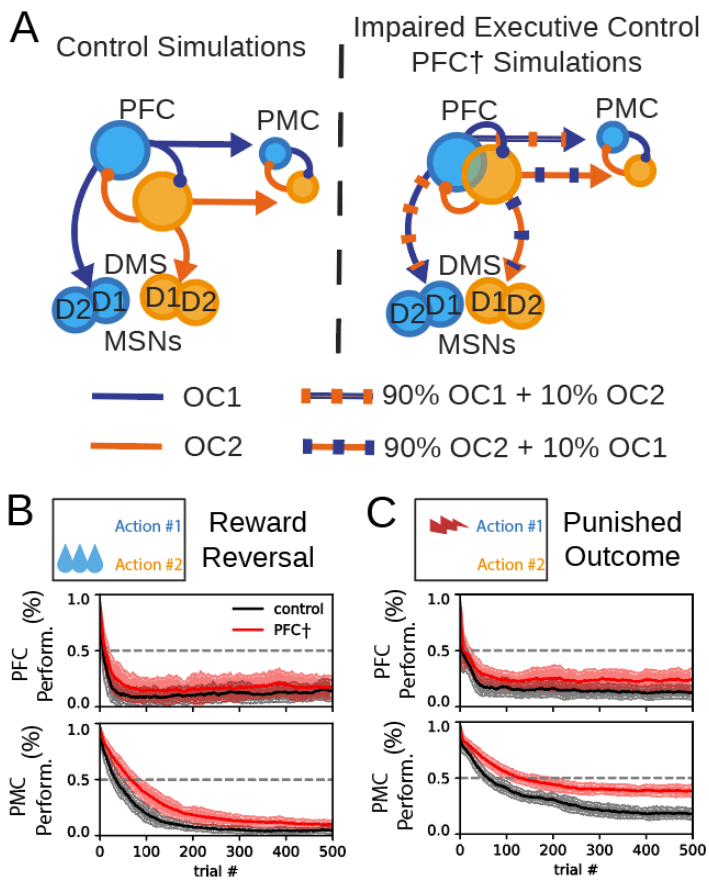

Figure 6. Impairment of PFC decreases learning performance.
(A). Illustration of impairment of PFC coding. The projections of the PFC into the DMS and the PMC contained mixed signals in $\mathrm{PFC}^{\dagger}$ behavioral sessions. OC1 and OC2 correspond to projections that represent the channels for 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 punished outcome task (black) and a punishment learning task with impaired PFC (red).

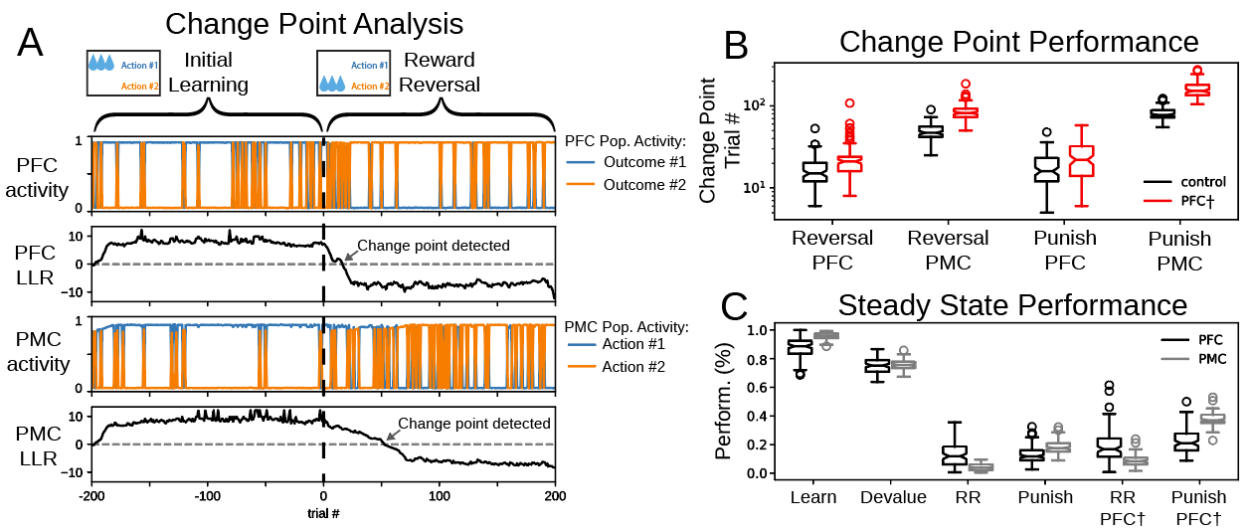

Figure 7. Analysis of model output across behavioral tasks

(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.