Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Neural reinforcement: re-entering and refining neural dynamics leading to desirable outcomes

Vivek R Athalye[1], Jose M Carmena[2] and Rui M Costa[1]

How do organisms learn to do again, on-demand, a behavior that led to a desirable outcome? Dopamine-dependent cortico-striatal plasticity provides a framework for learning behavior's value, but it is less clear how it enables the brain to re-enter desired behaviors and refine them over time. Reinforcing behavior is achieved by re-entering and refining the neural patterns that produce it. We review studies using brain-machine interfaces which reveal that reinforcing cortical population activity requires cortico-basal ganglia circuits. Then, we propose a formal framework for how reinforcement in cortico-basal ganglia circuits acts on the neural dynamics of cortical populations. We propose two parallel mechanisms: i) fast reinforcement which selects the inputs that permit the re-entrance of the particular cortical population dynamics which naturally produced the desired behavior, and ii) slower reinforcement which leads to refinement of cortical population dynamics and more reliable production of neural trajectories driving skillful behavior on-demand.

Addresses
[1] Zuckerman Mind Brain Behavior Institute, Departments of Neuroscience and Neurology, Columbia University, New York, NY, USA
[2] Helen Wills Neuroscience Institute, Department of Electrical Engineering and Computer Sciences, University of California-Berkeley, Berkeley, CA, USA

Corresponding author: Costa, Rui M (rc3031@columbia.edu)

Current Opinion in Neurobiology 2020, **60**:145–154

This review comes from a themed issue on **Neurobiology of behavior**

Edited by **Michael Brecht** and **Richard Mooney**

https://doi.org/10.1016/j.conb.2019.11.023

Introduction: motor reinforcement is neural reinforcement

When organisms discover a desirable outcome after performing a particular movement, they increase the occurrence of that movement. This behavioral principle is Thorndike's Law of Effect [1]. How does an organism learn which movement led to the outcome, and importantly, how does it select again that movement at the next attempt? The problem of credit assignment and re-entrance is crucial for understanding learning [2]. This problem can become easier to solve if we consider that the animal does not need to exactly re-do the same movement that was reinforced, but instead can re-enter a set of behaviors that obeys similar task-constraints and increases the probability of obtaining an outcome. Indeed, early in training organisms seem to re-enter behaviors similar to those that led to the outcome from the repertoire of attempts, with variation from trial to trial [3–5]. With further reinforcement, they gradually reduce variability and refine movement to more directly and efficiently achieve the outcome [6,7].

Motor reinforcement depends on dopamine-dependent plasticity at cortico-striatal synapses [8–11]. During reinforcement, neural ensembles and activity patterns in cortico-basal ganglia circuits are gradually refined, following a similar pattern to the movement [2,8,12–16]. This is logical, as learning occurs because the neural patterns producing actions that lead to reinforcement are re-entered more frequently and more precisely. Just as the organism does not need to replicate the exact behavior that led to an outcome, the brain may not need to replicate the exact pattern of neural activity. First, there is degeneracy: many neural activity patterns can cause the same motor output because many neurons in several motor control centers possess parallel lines of influence to the spinal cord driving muscles [17,18]. Second, entering a set of neural patterns that leads to a set of behaviors that obeys similar task-constraints to the target increases the probability of obtaining an outcome. Thus, the brain can learn from reinforcement to produce neural population activity within a target activity set which produces behavior within a target behavior set. Hence, the motor reinforcement problem is a neural reinforcement problem. In this review, we will focus on discussing how the brain re-enters a target activity set in motor cortex neurons for two reasons. First, motor cortex can control movement directly through spinal cord projections. Second, and perhaps more importantly, motor cortex can coordinate many motor control centers in the brain including the basal ganglia, thalamus, midbrain, brainstem, and spinal circuits, and they are the only class of supraspinal neurons that possess these broadcasting projections [18]. We will present a perspective that defends that what is reinforced through plasticity at different time courses and locations in cortico-basal ganglia circuits [16] are the outcome-relevant population dynamics in the cortex, and that this plasticity leads to fast re-entrance and gradual refinement of cortical population

dynamics, which underlies the learning of what behavior is relevant and how to perform it.

The brain learns from reinforcement to re-enter and refine cortical population activity controlling a brain-machine interface

Studying the neural reinforcement principles underlying behavioral reinforcement is challenging because it is hard to identify and observe the exact neural populations controlling behavior, and hence to interpret the impact of changes in neural activity in particular neurons on the occurrence of behavior. Closed-loop brain-machine interfaces (BMIs) are a useful experimental tool because the experimenter defines the mapping from observed neural activity to behavior, thus defining the target activity set that triggers a desirable outcome, such as a reward.

The brain can consolidate activity patterns for BMI control much like it does for motor control. Early work in the 1970s found that activity of individual motor cortex neurons could be reinforced if given sensory feedback on the firing rate [19,20]. Initial research on BMIs optimized to decode information about natural movement found that through closed-loop experience of the consequence of neural activity on the BMI, subjects improved BMI control, decreased neural variability, and changed neural encoding of prosthetic movement [21–24].

More recent primate studies established that the brain can learn to re-enter motor cortex population activity to control a BMI [25]. With training the neural patterns for BMI control gradually stabilized and were readily recalled, similar to neural activity evolution during motor reinforcement. The large initial trial-to-trial neural and neuroprosthetic variability decreased, and consistent neural trajectories emerged resulting in fast and direct neuroprosthetic movements [25,26•]. This gave insight into how the brain learns to re-enter and then gradually refine neural activity and behavior. Initially, the brain learned by modulating both the neurons controlling the BMI (direct neurons) as well as surrounding cortical network (indirect neurons), but gradually, the brain reduced modulation of indirect neurons relative to direct neurons [27,28]. Investigating the fine-scale spatial structure of credit assignment, a closed-loop optical BMI approach using 2-photon calcium imaging showed that early in training mice modulated the activity of the target direct neurons but also neighboring indirect neurons in close spatial vicinity; as training progressed animals learned to mostly modulate the activity of direct neurons [29], and learning could even be localized to an individual direct neuron [30]. Additional primate BMI studies studying adaptation to decoder perturbations [31–33] found that even within the direct neuron population the brain learns to assign credit to the direct neurons specifically driving behavioral error.

Cortico-basal ganglia circuits are necessary for reinforcement of cortical activity

Many experiments show that motor reinforcement depends on plasticity at the glutamatergic inputs to striatum, namely cortico-striatal synapses [8–11]. The striatum receives glutamatergic input carrying sensory, motor, and cognitive information from across cortex, parts of thalamus, and limbic areas as well as dopaminergic input carrying reinforcement signals from the midbrain (ventral tegmental area VTA and substantia nigra pars compacta SNc). The output of the basal ganglia (GPi, SNr) modulates motor control by directly or indirectly inhibiting or disinhibiting downstream brainstem populations that can serve as command lines to spinal cord circuits [18]. Thus, the basal ganglia are poised to use outcome-related signals to learn a mapping from environmental and cognitive states into control of brainstem activity to produce behaviors that lead to desirable outcomes. However, basal ganglia output also directly inhibits/disinhibits the thalamus and hence cortex [34]. Is plasticity at glutamatergic synapses in striatum necessary to learn to produce patterns of cortical activity that lead to desirable outcome?

Closed-loop BMI experiments showed that indeed cortico-striatal plasticity is necessary for learning to re-enter target motor cortex activity [35•]. As rodents learned to produce a specific pattern of activity in motor cortex to obtain reward, dorsal striatum neurons developed target-predictive modulation of activity [35]. Direct neurons in cortex developed coherence with dorsal striatum spiking and local field potential, which was not seen with indirect neurons [36]. Importantly, mice lacking the essential NR1 subunit of the NMDA receptor specifically in striatal projection neurons, and hence with disrupted cortico-striatal plasticity, could not learn to re-enter a cortical pattern that leads to reward more frequently than chance. Thus, plasticity in the cortico-striatal network resulted in task-related striatum activity and credit assignment to specific cortical neurons and patterns driving behavior.

The ability to reinforce specific behaviors that lead to reward relies on the activity of midbrain dopamine neurons, which modulates cortico-striatal plasticity [37–39]. Dopamine neurons in the ventral tegmental area (VTA) are thought to encode a reward-prediction error which signals difference between the animal's predicted and received reward [40] and is useful for reinforcing behavior [3]. This phasic activity of VTA is sufficient to reinforce specific behaviors that precede it, as animals learn to reproduce target behaviors that trigger VTA self-stimulation [41–44].

But how is the specific behavior reinforced and re-entered more frequently? It seems logical that neural patterns that lead to phasic VTA activity would be reinforced, constituting a Neural Law of Effect [45•]. A rodent BMI experiment found evidence for this principle: mice

learned to re-enter more frequently a rare motor cortex activity pattern that triggers optogenetic VTA self-stimulation [45•]. The striatum receives input from most cortices, and it would be adaptive for reinforcement to act even in primary sensory cortices, which encode not only antecedent stimuli but also exhibit internally driven, task-relevant modulations [46]. An intriguing BMI experiment found that rats and mice can learn to re-enter a target activity set in primary visual cortex to receive reward [47]. Learning was prevented when neurons in the dorsomedial striatum (DMS), which receives input from visual cortex, were optogenetically inhibited, but not during inhibition of nearby neurons in the dorsolateral striatum. Interestingly, after learning, DMS inhibition did not affect target activity re-entrance, indicating that the striatum is crucial for learning to re-enter cortical patterns that lead to desired outcomes, but not necessarily to execute that entrance after learning.

The closed-loop BMI experiments discussed above demonstrated that animals can learn to produce cortical patterns that lead to desired outcomes, such as they can produce behaviors, and that cortico-basal ganglia loops and dopamine are crucial for this reinforcement. Here, we propose a framework for how neural reinforcement acts on the dynamics of cortical populations. We propose two parallel mechanisms for how the brain uses cortico-basal ganglia circuits to re-enter and refine cortical population activity leading to desirable outcome: fast reinforcement to re-enter particular cortical population dynamics which naturally produces target activity resulting in variable but successful behavior, and slower reinforcement of cortical population dynamics which leads to refinement of neural ensembles and reliable production of neural trajectories driving skillful behavior on-demand.

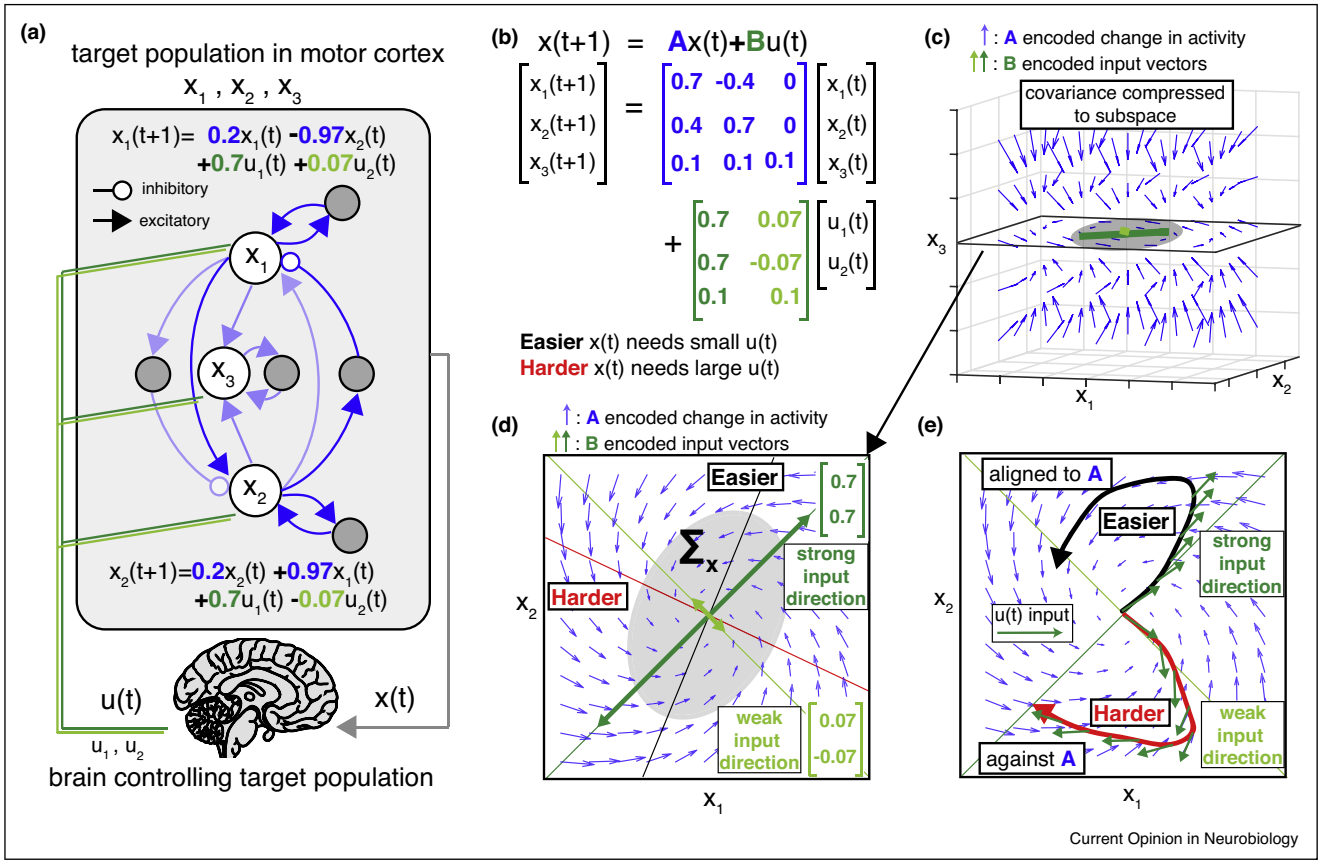
Re-entering neural activity to achieve desired outcome: reinforcement learning of neural dynamics and control

Let us now revisit our main question of how organisms learn to re-enter on-demand neural population activity that led to a desirable outcome. To address this question we propose to first consider how a particular pattern of neural activity is entered in the first place using a framework of neural dynamical systems. Neural dynamics has been critical for understanding motor cortex activity and its relationship to movement [48–53•]. The input connectivity and within-population connectivity determine the **neural dynamics** of a population of neurons; that is, the rules which dictate how population activity transitions across time and neurons (Figure 1a,b), as a function of input and connectivity, and defines activity dimensions (Figure 1c,d) and trajectories which are easier/harder to be entered (Figure 1e). Population dynamics render-specific transitions of activity natural and easy, and others more difficult to achieve. For example, let us model neural population activity as a linear dynamical system (Figure 1a,b), which has well-developed theory: $x(t+1) = Ax(t) + Bu(t)$, where $x(t)$ is a vector containing the activity of each neuron in the population, $u(t)$ is a vector of inputs to the population, the A matrix captures dynamics within the neural population, and B is the input matrix which maps input signals into population activity. While both $Ax(t)$ and $Bu(t)$ drive $x(t+1)$, there is a crucial distinction: dynamics $Ax(t)$ should be general across different goals, while input $u(t)$ can vary for different goals.

This can give us a conceptual basis to understand how easy or difficult it is for the brain to enter a particular population activity pattern, as recently reviewed [54]. The covariance (in steady-state) of motor cortex activity Σ_x defines and orders dimensions of neural activity by how much activity prefers to enter them (Figure 1c,d). It relates to the interaction of A, B, and is the solution to the Lyapunov equation $\Sigma_x = A\Sigma_x A^T + B\Sigma_u B^T$. Neural covariance relates to the minimum energy (or effort) of u needed to reach a target neural activity pattern x^* given by: $E_{min} = x^{*T}\Sigma_x^{-1}x^*$. (This expression assumes there's no time limit on reaching x^* and is calculated with $\Sigma_u = I$ to encode that each input can be equally and independently used.) Thus, from the control perspective, achieving a target on an activity dimension v with large neural covariance $\sigma(v) = v^T\Sigma_x v$ requires small external input (and a target on a low-covariance dimension requires large external input). If $\sigma(v) = 0$, activity can not at all be reached along dimension v. The dimensions which have large σ are an activity subspace which has been termed the population's 'intrinsic manifold' [55,56•].

Now we can consider, given a dynamics view of how population activity is entered, how is activity re-entered? We postulated that the brain does not learn by precisely replicating initial successful activity. Reinforcement learning theory presents the setting of an agent interacting with its environment, learning a policy $\pi_\theta a(t) \mid s(t))$ parameterized by θ which maps states s to a distribution over actions a in order to maximize rewards r [3]. We propose a model, in which the brain learns to adjust parameters of control and population dynamics in parallel to maximize reward. This makes the target population activity easier to achieve (i.e. less input u is needed), which makes the target activity more likely to be re-entered even with noisy input. We define a probability distribution describing the brain's interaction with the environment as follows: population activity $x(t)$ and behavior $y(t)$ update behavior $y(t+1)$ and reward $r(t+1)$ with parameters b unknown to the brain: $p_b(y(t+1), r(t+1) \mid x(t), y(t))$. Let the neural population be feedback-controlled with parameter w via $u(t) = g_w(x(t), y(t))$. (Note, we can model feedback being a function of only neural population activity or only of behavior.) The brain modifies initial parameters

Figure 1



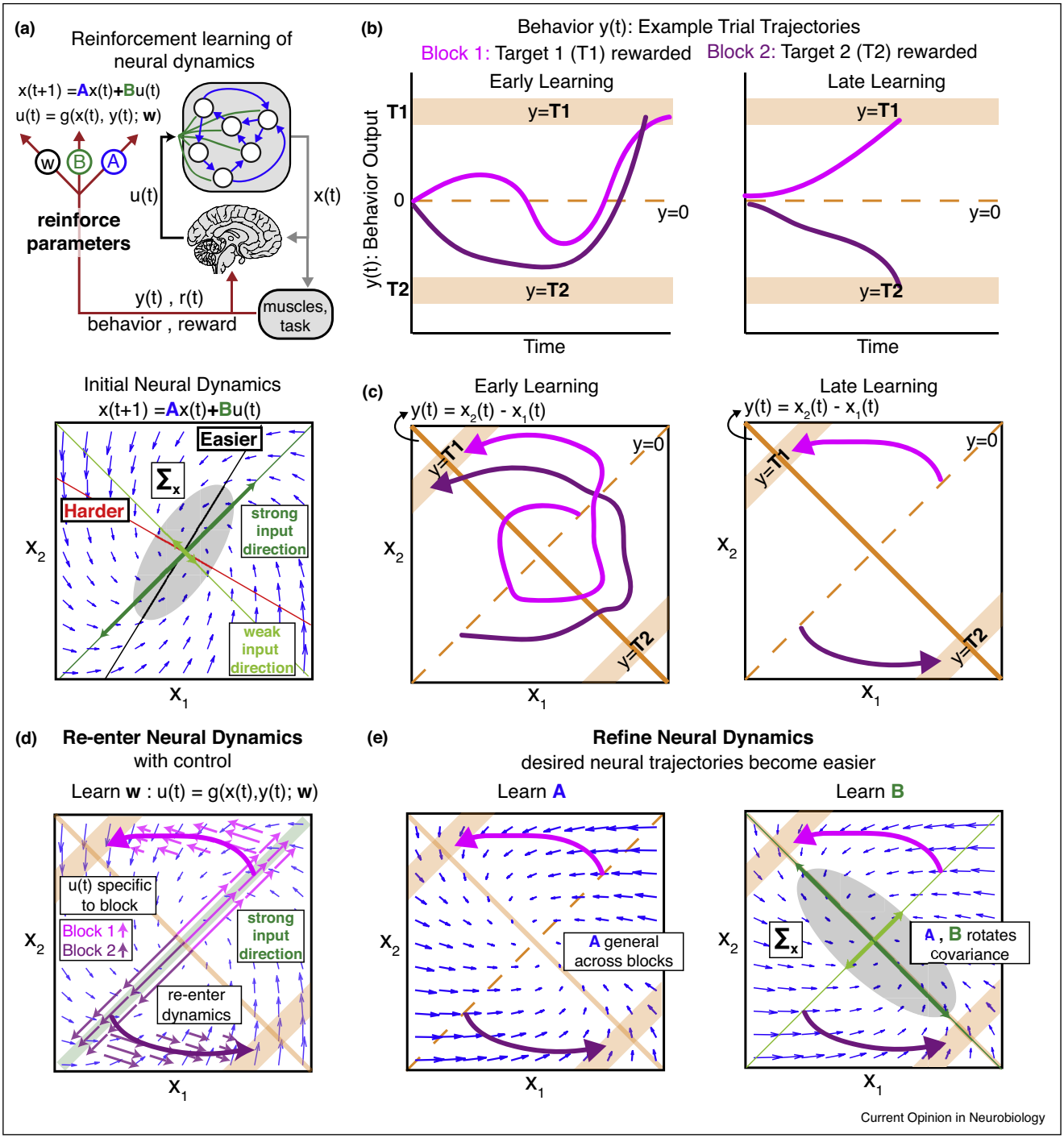
Entrance of target population activity through control of neural dynamics.
(a) A neural dynamics view of how activity is entered in a motor cortex population. Here is an illustrative example of the brain sending input u(t) to control activity x(t) of a small population of recurrently connected neurons. The network visualization motivates the relationship between physical connections and activity dynamics. Strong cortical connections are bold blue, weak are light blue. Neural dynamics are modeled with linear equations. **(b)** Linear dynamical system modeling of the pairwise interactions between neurons and input. **(c)** The effect of **A** matrix dynamics is shown by blue arrows, which show the direction neural activity prefers to evolve at each point in neural activity space, where each axis is one neuron's activity. This 3-neuron example reveals strong covariance in the x1–x2 subspace. In general, the covariance subspace captures activity modulations broadly across observed neurons, but for the sake of transparency, this example exhibits most covariance between neuron 1 and 2. **(d)** We zoom in on the plane of prominent neural covariance, where each axis is one neuron's activity. Dark green vector shows the input direction of the **B** matrix which has most gain and thus requires least u(t) input to affect neural activity, and the light green input vector has the least gain and thus requires the most input to affect neural activity. The neural covariance is shown in the gray ellipsoid, which reveals activity directions which are easier to move along (labeled with green 'Easier'), and directions which are harder to move along (labeled with red 'Hard'). Note that the black primary axis of covariance is not aligned with the strong input direction because the **A** matrix dynamics performs counter-clockwise rotation (visualized with blue arrows at each point of activity space). **(e)** Two example neural activity trajectories are shown. The black trajectory is easy to achieve, because it initially moves along the strong input direction, and then flows along neural dynamics. The red trajectory is difficult because it initially moves along the weak input direction, and then moves opposite to the direction of neural dynamics.

$\theta = A, B, w$ (Figure 2a) to learn reward-obtaining behavior (Figure 2b) through neural activity $x(t) \sim \pi_\theta x(t) | x(t-1), y(t-1))$ (Figure 2c). We will call A, B parameters of neural dynamics, as they map directly into neural activity space, and call w parameters of control. Learning w can select particular dimensions within the subspace defined by A, B, and controls how neural activity transitions in time within those dimensions (Figure 2d). Learning A, B can modify the subspace neural activity preferentially occupies to be oriented to target activity (Figure 2e). Learning B changes the directions and gain by which input drives activity. Learning A modifies how neural activity acts on itself, which can be visualized as a flow field in neural activity space.

Studying reinforcement learning of neural dynamics with brain-machine interfaces

Closed-loop BMIs can be used to test how the brain learns activity patterns which have a particular relationship to pre-existing neural dynamics. A relevant study hypothesized that it is easier for the brain to learn neural patterns within its pre-existing covariance [56•,57] (Figure 3a), which

Figure 2



Re-entrance of target population activity through reinforcement learning of neural dynamics.
We illustrate how re-entrance of target neural activity through reinforcement learning can underlie motor learning. **(a)** Reinforcement learning of neural activity may operate on parameters of neural dynamics. We identify three parameters: 1) **w** which maps cortical activity x(t) and behavior y (t) to input u(t) to the target cortical population, 2) **B** which are weights mapping input u(t) to cortical activity x(t), and 3) **A** which are cortical dynamics weights determining rules of how cortical activity evolves as a function of its past. For illustration, we introduce a two-neuron population with depicted initial dynamics. Each axis is one neuron's firing rate. **A** is represented by blue flow fields, **B** is represented by green arrows; dark green arrow shows the strong input direction, and the light green arrow shows the weak input direction. The primary axes of the gray covariance ellipse indicate the directions where activity is easy and hard to reach. **(b)** We introduce an illustrative motor task, where one dimensional behavior y(t) triggers reinforcement. In Block 1, Target 1 (T1) is rewarded, and in Block 2, Target 2 (T2) is rewarded. One example trajectory from Block 1 and Block 2 are shown. In early learning, an error trial to T1 performed in block 2 is shown. In late learning, behavior is more successful, less

captures activity dimensions which neural dynamics preferentially occupies, than patterns outside. This mathematically defined hypothesis was translated into a BMI learning experiment where some decoders were designed to align with neural covariance and others to misalign. Subjects could readily learn within a day decoders aligned with neural covariance but not misaligned decoders (Figure 3b,c), providing compelling evidence for their hypothesis, suggesting that if the brain has pre-existing cortical population dynamics (A, B) that are relevant to solve a task, it is faster to learn to control those dynamics (w) than to modify the dynamics.

This study focused on BMI-experienced subjects learning within one day. However, we discussed that the brain can learn to control difficult decoders, that is, re-enter difficult patterns, over many days [25,27,58] (Figure 3d). In early learning, indirect neurons were modulated along with direct neurons [27–29], suggesting that the brain initially learns by re-entering dynamics of a population larger than the direct neurons. Gradual credit assignment to direct neurons implied the brain is able to reinforce the variation direct neurons exhibit which is largely independent of other neurons and leads to reward [27–29]. To study the balance of shared and individual (independent) neural variance during neural re-entrance, we analyzed neural covariance during long term *de novo* BMI learning [26•,45•]. In this setting, the neural population exhibited very little shared variance in early learning (Figure 3e). With training, large initial trial-to-trial variability of individual neurons and the cursor decreased. Low-dimensional shared variance increased, and the subspace of shared variance rotated and stabilized, aligning to the decoder and making shared variation more efficient in driving the cursor (Figure 3f). Finally, consistent neural trajectories within the shared subspace emerged which produced skillful control (Figure 3f). Crucially, we found this same evolution of individual and shared variance in mice learning to re-enter specific cortical patterns in order to receive optogenetic VTA self-stimulation [45•], establishing a role for dopamine and the basal ganglia in the reinforcement of cortical dynamics. Reinforcement of high dimensional neural exploration permits learning novel neural patterns, as found in recent work studying long-term learning of difficult decoders misaligned with initial neural covariance [58].

These studies support the hypothesis that learning control of cortical dynamics is faster while modifying cortical population dynamics is slower (Figure 3a,d). Future BMI experiments can further test the relationship between neural dynamics and reinforcement. One relevant study paved the way for designing BMI experiments to test properties of neural dynamics [53•] and demonstrated the power of single trial predictions of neural dynamics models by designing a high-performance closed-loop BMI.

Cortico-basal ganglia circuits for neural reinforcement through re-entrance and refinement of cortical population dynamics

We propose that the cortico-basal ganglia circuit is crucial for neural reinforcement by *i)* initially facilitating re-entering of the right cortical population dynamics and *ii)* gradually facilitating the refinement of cortical dynamics to achieve target activity more directly and reliably (Figure 4).

i) Re-entrance of cortical population dynamics

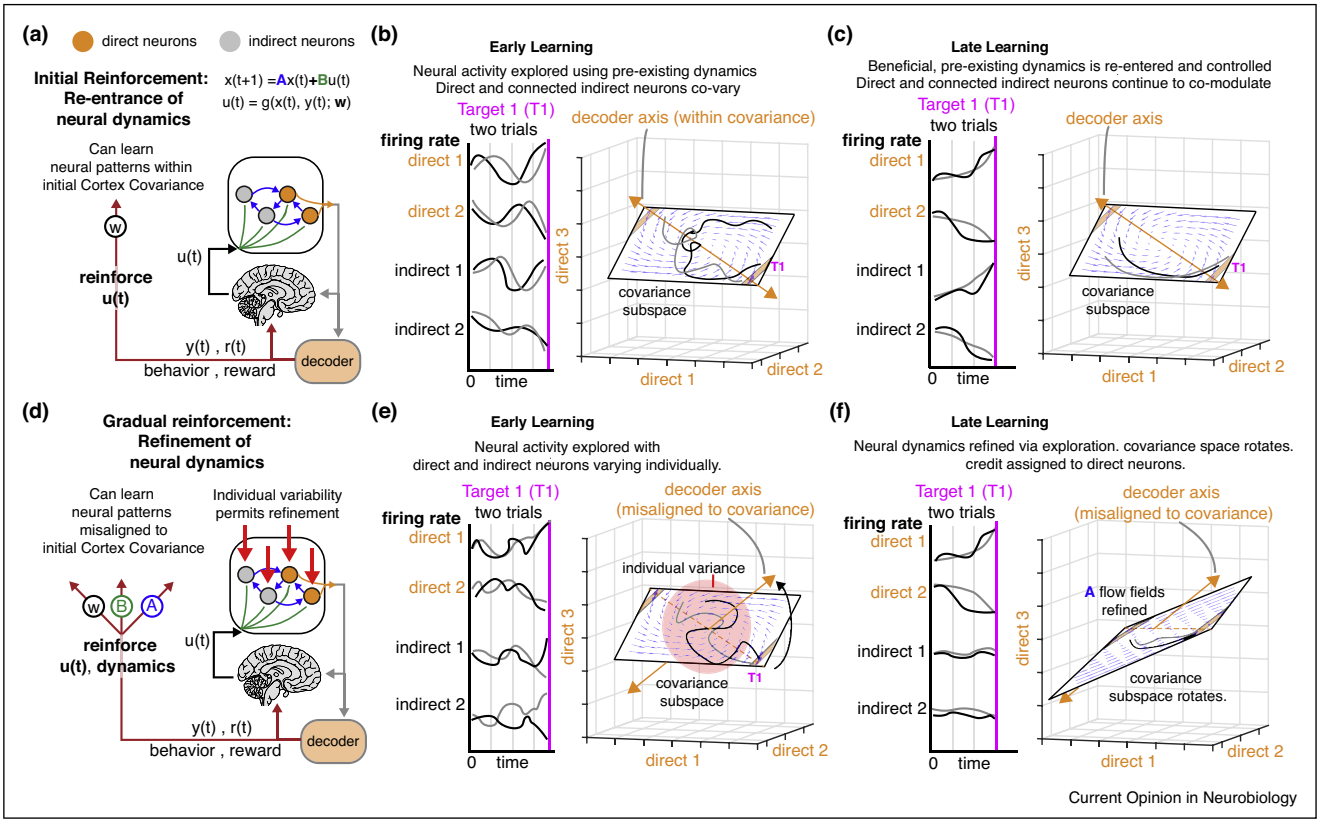
When cortex first enters target activity, midbrain dopamine neurons send a reward prediction error signal which projects to striatum, tagging the preceding neural activity as special, and triggering the process of plasticity crucial for reinforcement. The set of cortical neurons which underlie the first success are a subset of a target population whose dynamics and output connections make it natural for generating target activity. Further, many different population activity patterns compose the target activity set, as previously discussed. Thus, the brain must find one of many inputs controlling the target population to re-enter target activity. How can the brain do this?

The cortico-basal ganglia circuit is topographically organized in re-entrant loops [59,60], such that the basal ganglia can produce outputs which feedback to a targeted population in cortex. In particular, when dopamine reinforces synapses from the 'initial success cortical neurons' to striatum, the basal ganglia is organized such that activity propagating from synaptically modified striatum neurons through the basal ganglia and thalamus will re-enter a population of cortical neurons containing the 'initial successful cortical neurons'.

We propose the basal ganglia selects task-relevant dimensions within the target population's pre-existing covariance, by learning low-dimensional common inputs to the target

(Figure 2 Legend Continued) variable, faster, and more direct. **(c)** We show the neural activity producing the behavior in B). Behavior is produced by the difference of the two neurons' activity $y(t) = x_2(t) - x_1(t)$, as shown by the dark orange line. Visually, the projection of neural activity onto the dark orange line results in behavior y(t). The orange target rectangles demarcate target activity which neural trajectories must enter to achieve target behavior. The dotted orange line is the axis that produces behavior $y = 0$. **(d)** The brain can learn **w** to produce u(t), which allows the brain to re-enter and control useful neural dynamics. This results in goal-specific inputs to the target neural population, which can be visualized as goal-specific arrows in neural activity space. Inputs must still propagate with dynamics parameters **A**, **B** as shown in panel C). **(e)** The brain can learn **A**, **B**, which refines neural dynamics. Learning **A** changes the flow fields which facilitates entering target neural activity and generalizes across blocks. Note that the flow fields make sample-by-sample predictions about how neural activity transitions in time, and can't be estimated in general with just two trials. Learning **B** changes the directions on which inputs can move activity. The combination of learning **A**, **B** can rotate neural covariance. In contrast, learning **w** without learning **A**, **B** merely selects activity from covariance permitted by **A**, **B**.

Figure 3



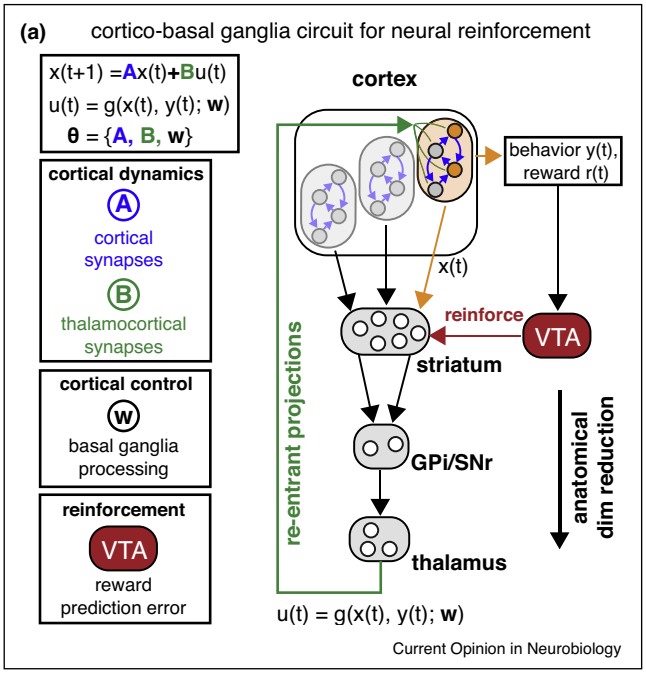
Interpretation of BMI studies for reinforcement of neural dynamics.
(a) BMI learning experiments have revealed principles of how neural activity is learned. Reference [56•] designed BMI experiments where decoders are either aligned or misaligned with motor cortex covariance, and found that the brain more easily learns motor cortex patterns within its pre-existing covariance. The study supports the idea that the brain performs initial reinforcement by re-entering and controlling pre-existing neural dynamics. **(b)** Early learning activity is explored using pre-existing neural dynamics, which lies prominently in the covariance space or 'intrinsic manifold'. We predict this covariance would extend to other connected neurons in the cortical network which are not decoded for the BMI. **(c)** Late learning results in re-entrance and control of beneficial, pre-existing neural dynamics, which corresponds to selection of neural activity from cortical covariance. We predict connected indirect neurons continue to co-modulate with the task. **(d)** Other BMI experiments found that with long-term training, neural covariance can be modified [26,45•], and behavior-driving neurons are assigned credit with learning [27–29]. We interpret these results to mean that cortical dynamics can be refined with long-term training. This process depends on direct neurons varying independently from indirect neurons. **(e)** Subjects learn a decoder which is misaligned to initial cortex covariance. In early learning, direct and indirect neurons modulate to the task, but neurons show prominent individual variation which explores activity beyond pre-existing dynamics. **(f)** In late learning over days of training, direct neurons modulate to achieve target activity while indirect neurons reduce modulation, revealing credit assignment to direct neurons. Within the direct neuron population, covariance rotates to align with the decoder axis, making achieving target neural activity more efficient. Consistent neural trajectories emerge producing skillful behavior. Likely, network connections are modified to refine neural dynamics, constructing improved flow fields for generating target activity.

population via thalamus. Dopamine-dependent cortico-striatal plasticity permits the striatum to rapidly learn task-relevant patterns in cortex which lead to reward [35•,36], and propagate them through basal ganglia output down to the brainstem, but also back to the cortex via thalamus. The basal ganglia constitutes a significant anatomical bottleneck from its cortical inputs to its GPi/SNr outputs [61]. In the rat, ~20 million cortical neurons (nearly all of cortex) projects to dorsal striatum [62], while SNr has ~25 000 neurons [63], and interconnectivity of output neurons reduces the dimensionality of its activity [64]. This low-dimensional output from GPi/SNr has a reduced parameter space compared to the activity space of the initial cortical pattern and projects to brainstem to disinhibit wanted movements/patterns, but also to thalamus and then to cortex. Computationally, these anatomical constraints may place the basal ganglia as a natural learner of the low-dimensional task-relevant dynamics in the cortex, and a selector of the appropriate input back to the cortex to re-enter these dynamics. Notably, this re-entrance does not require plasticity within the cortex.

ii) Refinement of cortical population dynamics

As the brain re-enters cortical population dynamics to produce variable but successful activity, cortical dynamics

Figure 4



Cortico -basal ganglia circuit for neural reinforcement through re-entrance and refinement of neural population dynamics. We highlight two properties. 1) The basal ganglia is positioned to learn low-dimensional inputs to cortex due to anatomical dimensionality reduction from striatum input to GPi/SNr output. 2) The basal ganglia can reinforce activity in particular cortical populations because it has parallel, re-entrant loops which feedback to particular neural populations. Illustrated is the particular loop which feeds back to a neural population containing neurons responsible for the target behavior. When cortex first enters target activity, VTA fires a dopaminergic reward prediction error signal, which strongly projects to cortico-striatal synapses and plays an important role in reinforcement. We hypothesize that learning **w**, a mapping from high-dimensional cortical activity and behavioral signals to low-dimensional cortical input u(t), may be responsible for initial reinforcement which re-enters pre-existing cortical dynamics. Gradually, cortical dynamics are refined: thalamocortical synapses can be learned to map input to cortical activity (**B** matrix), and cortical synapses can be learned to modify cortical dynamics (**A** matrix) and enable desired neural trajectories to be produced on-demand.

are gradually refined. Less task-relevant neurons and trajectories are entered less often while task-relevant activity is optimized to more directly achieve an outcome and is triggered more reliably. We propose that this process involves Hebbian-like, dopamine-dependent plasticity at cortico-striatal synapses [38], which continues to refine with continued learning [10] but also requires plasticity in the cortex [13,65–67] and depends on sleep [28,68]. In this scenario, basal ganglia reinforces inputs to the motor cortex which could lead to the repeated co-activation of particular cortical cells in lieu of others, leading to the refinement of connectivity and dynamics via Hebbian plasticity in the cortex. Refining dynamics depends on reinforcement acting on variation of task-relevant neurons which is independent of less task-relevant neurons[26•,45•]. Modeling work has shown that with cortical neurons exhibiting uncorrelated variance and a three-factor dopamine-dependent STDP learning rule, credit assignment is possible to BMI direct neurons [69]. This plasticity in the cortex could potentially be mediated by direct dopaminergic inputs to the cortex [70,71]. Dynamics refinement is likely also mediated by thalamocortical plasticity which has shown specificity to learning-related cortical cells [72].

These refined neural dynamics [48] will enable the neural population to reliably produce neural trajectories [13,26•] which directly achieve the target behavior and the desired outcome. We also consider that in subsequent phases of training, where extensive training lead to consolidated cortical dynamics and to asymptotic refinement, basal ganglia plasticity and activity may no longer be required for skillful re-entrance of specific cortical patterns [47,73].

Conclusion

In conclusion, we propose a framework, in which reinforcement learning operates on neural dynamics, and show that such a framework can lead to hypotheses of how the brain learns to re-enter desired neural activity at different timescales, likely involving different forms of plasticity at different nodes of the basal ganglia-thalamo-cortical circuitry, consequential for motor behavior as well as abstract thought. Future experiments will be needed to test further predictions about how neural population dynamics facilitate and constrain learning of activity and behavior.

Conflict of interest statement

Nothing declared.

Acknowledgements

'This work was supported by an N.I.H. postdoctoral fellowship (BRAIN F32) to V.R.A. and a U19 Brain Initiative grant (5U19NS104649) to R.M.C.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. Thorndike EL: **Animal intelligence: an experimental study of the associative processes in animals**. *Psychol Rev* 1898, **2**:1-107.
2. Costa RM: **A selectionist account of de novo action learning**. *Curr Opin Neurobiol* 2011, **21**:579-586.
3. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. MIT Press; 1998.
4. Tumer EC, Brainard MS: **Performance variability enables adaptive plasticity of "crystallized" adult birdsong**. *Nature* 2007, **450**:1240-1244.
5. Wu HG, Miyamoto YR, Castro LNG, Ölveczky BP, Smith MA: **Temporal structure of motor variability is dynamically regulated and predicts motor learning ability**. *Nat Neurosci* 2014, **17**:312-321.

6. Cohen RG, Sternad D: **Variability in motor learning: relocating, channeling and reducing noise**. *Exp Brain Res* 2009, **193**:69-83.

7. Shmuelof L, Krakauer JW, Mazzoni P: **How is a motor skill learned? Change and invariance at the levels of task success and trajectory control**. *J Neurophysiol* 2012, **108**:578-594.

8. Barnes TD, Kubota Y, Hu D, Jin DZ, Graybiel AM: **Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories**. *Nature* 2005, **437**:1158-1161.

9. Dang MT, Yokoi F, Yin HH, Lovinger DM, Wang Y, Li Y: **Disrupted motor learning and long-term synaptic plasticity in mice lacking NMDAR1 in the striatum**. *Proc Natl Acad Sci U S A* 2006, **103**:15254-15259.

10. Santos FJ, Oliveira RF, Jin X, Costa RM: **Corticostriatal dynamics encode the refinement of specific behavioral variability during skill learning**. *eLife* 2015, **4**:e09423.

11. Yin HH, Mulcare SP, Hilário MRF, Clouse E, Holloway T, Davis MI, Hansson AC, Lovinger DM, Costa RM: **Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill**. *Nat Neurosci* 2009, **12**:333-341.

12. Cao VY, Ye Y, Mastwal S, Ren M, Coon M, Liu Q, Costa RM, Wang KH: **Motor learning consolidates arc-expressing neuronal ensembles in secondary motor cortex**. *Neuron* 2015, **86**:1385-1392.

13. Peters AJ, Chen SX, Komiyama T: **Emergence of reproducible spatiotemporal activity during motor learning**. *Nature* 2014, **510**:263-267.

14. Ölveczky BP, Otchy TM, Goldberg JH, Aronov D, Fee MS: **Changes in the neural control of a complex motor sequence during learning**. *J Neurophysiol* 2011, **106**:386-397.

15. Sheng M, Lu D, Shen Z, ming, Poo M: **Emergence of stable striatal D1R and D2R neuronal ensembles with distinct firing sequence during motor learning**. *Proc Natl Acad Sci U S A* 2019, **166**:11038-11047.

16. Costa RM, Cohen D, Nicolelis MAL: **Differential corticostriatal plasticity during fast and slow motor skill learning in mice**. *Curr Biol* 2004, **14**:1124-1134.

17. Hennig JA, Golub MD, Lund PJ, Sadtler PT, Oby ER, Quick KM, Ryu SI, Tyler-Kabara EC, Batista AP, Yu BM *et al.*: **Constraints on neural redundancy**. *eLife* 2018, **7**:1-34.

18. Arber S, Costa RM: **Connecting neuronal circuits for movement**. *Science (80-)* 2018, **360**:1403-1404.

19. Fetz EE: **Operant conditioning of cortical unit activity**. *Science (80-)* 1969, **163**:955-958.

20. Fetz EE, Baker MA: **Operantly conditioned patterns on precentral unit activity and correlated responses in adjacent cells and contralateral muscles**. *J Neurophysiol* 1973, **36**:179-204.

21. Taylor DM, Tillery SIH, Schwartz AB: **Direct cortical control of 3D neuroprosthetic devices**. *Science (80-)* 2002, **296**:1829-1832.

22. Carmena JM, Lebedev MA, Crist RE, O'Doherty JE, Santucci DM, Dimitrov DF, Patil PG, Henriquez CS, Nicolelis MAL: **Learning to control a brain-machine interface for reaching and grasping by primates**. *PLoS Biol* 2003, **1**:193-208.

23. Musallam S, Corneil BD, Greger B, Scherberger H, Andersen RA: **Cognitive control signals for neural prosthetics**. *Science (80-)* 2004, **305**:258-263.

24. Zacksenhouse M, Lebedev MA, Carmena JM, O'Doherty JE, Henriquez C, Nicolelis MAL: **Cortical modulations increase in early sessions with brain-machine interface**. *PLoS One* 2007, **2**:e619.

25. Ganguly K, Carmena JM: **Emergence of a stable cortical map for neuroprosthetic control**. *PLoS Biol* 2009, **7**.

26. • Athalye VR, Ganguly K, Costa RM, Carmena JM: **Emergence of coordinated neural dynamics underlies neuroprosthetic learning and skillful control**. *Neuron* 2017, **93**:955-970

Authors showed that long-term learning of difficult cortical patterns involves strengthening neural covariance, rotating it away from its initial subspace, and aligning it to task-relevant neural dimensions, making it more efficient in driving the BMI. Consistent neural trajectories emerged within the covariance subspace which produced skillful control. Initial exploration was accomplished with neural variability private to each neuron, which authors suggest enables credit assignment to the neurons controlling the BMI and rotation of covariance away from its initial subspace. These results suggest that long-term learning of difficult neural patterns relies on gradual reinforcement of neural population dynamics.

27. Ganguly K, Dimitrov DF, Wallis JD, Carmena JM: **Reversible large-scale modification of cortical networks during neuroprosthetic control**. *Nat Neurosci* 2011, **14**:662-667.

28. Gulati T, Guo L, Ramanathan DS, Bodepudi A, Ganguly K: **Neural reactivations during sleep determine network credit assignment**. *Nat Neurosci* 2017, **2017**:3-5.

29. Clancy KB, Koralek AC, Costa RM, Feldman DE, Carmena JM: **Volitional modulation of optically recorded calcium signals during neuroprosthetic learning**. *Nat Neurosci* 2014, **17**:807-809.

30. Prsa M, Galinanes GL, Huber D: **Rapid integration of artificial sensory feedback during operant conditioning of motor cortex neurons**. *Neuron* 2017, **93**:929-939.e6.

31. Chase SM, Kass RE, Schwartz AB: **Behavioral and neural correlates of visuomotor adaptation observed through a brain-computer interface in primary motor cortex**. *J Neurophysiol* 2012, **108**:624-644.

32. Jarosiewicz B, Chase SM, Fraser GW, Velliste M, Kass RE, Schwartz AB: **Functional network reorganization during learning in a brain-computer interface paradigm**. *Proc Natl Acad Sci U S A* 2008, **105**:19486-19491.

33. Zhou X, Tien RN, Ravikumar S, Chase SM: **Distinct types of neural reorganization during long-term learning**. *J Neurophysiol* 2019, **121**:1329-1341.

34. Oldenburg IA, Sabatini BL: **Antagonistic but not symmetric regulation of primary motor cortex by basal ganglia direct and indirect pathways**. *Neuron* 2015, **86**:1174-1181.

35. • Koralek AC, Jin X, Long IIJD, Costa RM, Carmena JM: **Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills**. *Nature* 2012, **483**:331-335

Authors showed that learning to operantly control cortical population activity requires cortico-striatal plasticity, linking cortico-basal ganglia circuits to cortical activity reinforcement. These experiments were the first to test and establish learning of goal-directed neuroprosthetic control and found striatum activity developed goal-predictive modulation with learning.

36. Koralek AC, Costa RM, Carmena JM: **Temporally precise cell-specific coherence develops in corticostriatal networks during learning**. *Neuron* 2013, **79**:865-872.

37. Reynolds JNJ, Hyland BI, Wickens JR: **A cellular mechanism of reward-related learning**. *Nature* 2001, **413**:67-70.

38. Shen W, Flajolet M, Greengard P, Surmeier DJ: **Dichotomous dopaminergic control of striatal synaptic plasticity**. *Science (80-)* 2008, **321**:848-851.

39. Yagishita S, Hayashi-Takagi A, Ellis-Davies GCR, Urakubo H, Ishii S, Kasai H: **A critical time window for dopamine actions on the structural plasticity of dendritic spines**. *Science (80-)* 2014, **345**:1616-1620.

40. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward**. *Science (80-)* 1997, **275**:1593-1599.

41. Olds J, Milner P: **Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain**. *J Comp Physiol Psychol* 1954, **47**:419-427.

42. Corbett D, Wise RA: **Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbraine: a moveable electrode mapping study**. *Brain Res* 1980, **185**:1-15.

43. Tsai H-C, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K: **Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning**. *Science (80-)* 2009, **324**:1080-1085.

44. Witten IB, Steinberg EE, Lee SY, Davidson TJ, Zalocusky KA, Brodsky M, Yizhar O, Cho SL, Gong S, Ramakrishnan C *et al.*:

Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron* 2011, **72**:721-733.

45. • Athalye VR, Santos FJ, Carmena JM, Costa RM: **Evidence for a neural law of effect**. *Science (80-)* 2018, **359**:1024-1029
Authors showed that mice learn to re-enter cortical neural activity which triggers optogenetic dopaminergic VTA stimulation. Dopamine-dependent reinforcement resulted in strengthening neural covariance and aligning it to task-relevant neural dimensions. These results link dopaminergic reinforcement to modification of cortical population dynamics.

46. Makino H, Komiyama T: **Learning enhances the relative impact of top-down processing in the visual cortex**. *Nat Neurosci* 2015, **18**:1116-1122.

47. Neely RM, Koralek AC, Athalye VR, Costa RM, Carmena JM: **Volitional modulation of primary visual cortex activity requires the basal ganglia**. *Neuron* 2018, **97**:1356-1368.

48. Shenoy KV, Sahani M, Churchland MM: **Cortical control of arm movements: a dynamical systems perspective**. *Annu Rev Neurosci* 2013, **36**:337-359.

49. Kaufman MT, Churchland MM, Ryu SI, Shenoy KV: **Cortical activity in the null space: permitting preparation without movement**. *Nat Neurosci* 2014, **17**:440-448.

50. Russo AA, Bittner SR, Perkins SM, Seely JS, London BM, Lara AH, Miri A, Marshall NJ, Kohn A, Jessell TM *et al.*: **Motor cortex embeds muscle-like commands in an untangled population response**. *Neuron* 2018, **97**:953-966.e8.

51. Churchland MM, Cunningham JP, Kaufman MT, Ryu SI, Shenoy KV: **Cortical preparatory activity: representation of movement or first cog in a dynamical machine?** *Neuron* 2010, **68**:387-400.

52. Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV: **Neural population dynamics during reaching**. *Nature* 2012, **487**:1-20.

53. • Kao JC, Nuyujukian P, Ryu SI, Churchland MM, Cunningham JP, Shenoy KV: **Single-trial dynamics of motor cortex and their applications to brain-machine interfaces**. *Nat Commun* 2015, **6**:7759
Authors demonstrated the power of statistically modeling cortical population dynamics by designing a high performance BMI. This work shows that neural dynamics makes single trial predictions of how neural activity evolves in time and opens the door for designing BMI experiments to probe which activity patterns are naturally produced by observed neural population dynamics.

54. Kao T, Hennequin G: **Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics**. *Curr Opin Neurobiol* 2019, **58**:122-129.

55. Gallego JA, Perich MG, Miller LE, Solla SA: **Neural manifolds for the control of movement**. *Neuron* 2017, **94**:978-984.

56. • Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, Yu BM, Batista AP: **Neural constraints on learning**. *Nature* 2014, **512**:423-426
Authors designed BMI learning experiments to show that it is easier for the brain to learn neural patterns within its pre-existing covariance, which captures activity dimensions which neural dynamics preferentially occupies, than patterns outside. These results suggest that if the brain possesses pre-existing cortical population dynamics that are relevant to solve a task, it is faster to learn control of those dynamics from reinforcement than to modify the dynamics.

57. Golub MD, Sadtler PT, Oby ER, Quick KM, Ryu SI, Tyler-Kabara EC, Batista AP, Chase SM, Yu BM: **Learning by neural reassociation/631/378/1595/631/378/2629 article**. *Nat Neurosci* 2018, **21**:607-616.

58. Oby ER, Golub MD, Hennig JA, Degenhart AD, Tyler-Kabara EC, Yu BM, Chase SM, Batista AP: **New neural activity patterns emerge with long-term learning**. *Proc Natl Acad Sci U S A* 2019, **116**:15210-15215.

59. Alexander GE, DeLong MR, Strick PL: **Parallel organization of functionally segregated circuits linking basal ganglia and cortex**. *Annu Rev Neurosci* 1986, **9**:357-381.

60. Redgrave P, Vautrelle N, Reynolds JNJ: **Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement**. *Neuroscience* 2011, **198**:138-151.

61. Dudman JT, Krakauer JW: **The basal ganglia: from motor commands to the control of vigor**. *Curr Opin Neurobiol* 2016, **37**:158-166.

62. Zheng T, Wilson CJ: **Corticostriatal combinatorics: the implications of corticostriatal axonal arborizations**. *J Neurophysiol* 2002, **87**:1007-1017.

63. Oorschot DE: **Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: a stereological study using the cavalieri and optical disector methods**. *J Comp Neurol* 1996, **366**:580-599.

64. Brown J, Pan WX, Dudman JT: **The inhibitory microcircuit of the substantia nigra provides feedback gain control of the basal ganglia output**. *eLife* 2014, **2014**:1-25.

65. Rioult-Pedotti MS, Friedman D, Donoghue JP: **Learning-induced LTP in neocortex**. *Science (80-)* 2000, **290**:533-536.

66. Rioult-Pedotti M-S, Friedman D, Hess G, Donoghue JP: **Strengthening of horizontal cortical connections following skill learning**. *Nat Neurosci* 1998, **1**:230-234.

67. Hayashi-Takagi A, Yagishita S, Nakamura M, Shirai F, Wu YI, Loshbaugh AL, Kuhlman B, Hahn KM, Kasai H: **Labelling and optical erasure of synaptic memory traces in the motor cortex**. *Nature* 2015, **525**:333-338.

68. Yang G, Sau Wan Lai C, Cichon J, Ma L, Li W, Gan W: **Sleep promotes branch-specific formation of dendritic spines after learning**. *Science (80-)* 2014, **344**:1173-1178.

69. Legenstein R, Chase SM, Schwartz AB, Maass W: **A reward-modulated Hebbian learning rule can explain experimentally observed network reorganization in a brain control task**. *J Neurosci* 2010, **30**:8400-8410.

70. Hosp JA, Pekanovic A, Rioult-Pedotti MS, Luft AR: **Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning**. *J Neurosci* 2011, **31**:2481-2487.

71. Guo L, Xiong H, Kim J-I, Wu Y-W, Lalchandani RR, Cui Y, Shu Y, Xu T, Ding JB: **Dynamic rewiring of neural circuits in the motor cortex in mouse models of Parkinson's disease**. *Nat Neurosci* 2015, **18**:1299-1309.

72. Biane JS, Takashima Y, Scanziani M, Conner JM, Tuszynski MH: **Thalamocortical projections onto behaviorally relevant neurons exhibit plasticity during adult motor learning**. *Neuron* 2016, **89**:1173-1179.

73. Makino H, Hwang EJ, Hedrick NG, Komiyama T: **Circuit mechanisms of sensorimotor learning**. *Neuron* 2016, **92**:705-721.