Learning to select actions shapes recurrent dynamics in the corticostriatal system

Christian D. Márton[1,2*¤], Simon R. Schultz[1], Bruno B. Averbeck[2]

1 Centre for Neurotechnology & Department of Bioengineering, Imperial College London, London, SW7 2AZ, U.K.

2 Laboratory of Neuropsychology, Section on Learning and Decision Making, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

¤Current Address: Laboratory of Neuropsychology, Section on Learning and Decision Making, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

* cdmarton@gmail.com

Abstract

Learning to select appropriate actions based on their values is a fundamentally human task which draws on the corticostriatal system in the brain. The prefrontal cortex (PFC) and dorsal striatum (dSTR) within this system are key in learning complex behaviors, and approaches from dynamical systems theory have allowed insight into how neural networks represent these behaviors. Yet, how learning itself is represented in top-down signals remains unknown. We hypothesized that learning is expressed in latent neural population dynamics. Therefore, we built a joint recurrent network model of the corticostriatal system and trained it on a complex learning task which involved executing the correct three-movement sequence. This system consisted of a striatal component which encoded action values and a prefrontal component which selected appropriate actions. After training, this system is able to autonomously predict value and select actions with the same performance as the animals trained on this task. We found that model representations mirrored those obtained from neural recordings in two macaques trained on the same task. We found that learning drove sequence-representations further apart from each other in latent space, both in our model and in the neural data. Our model revealed that learning proceeds by increasing the distance between sequence-specific fixed point regions and, hence, makes it more likely to select the appropriate action sequence. We also found that PFC sequence representations were more structured than dSTR representations. Altogether, we used a joint recurrent network model of the corticostriatal system together with neural recordings from the same regions to uncover the first evidence of how learning is expressed in top-down neural population signals within the corticostriatal system in the brain.

1 Introduction

Human and nonhuman primates are capable of complex, flexible and adaptable behavior. This ability relies on an equally complex interplay of regions in the brain [Seo et al., 2012, Verschure et al., 2014, Domenech and Koechlin, 2015]. Adaptable behavior requires predicting the values of choices, executing actions on the basis of those predictions, and updating predictions following the rewarding or non-rewarding outcomes of choices. Reinforcement learning (RL) in which rewarded actions are being reinforced is one framework through which to enable adaptable behavior. The power of this framework to drive complex behavior [**?**, Wang et al., 2018] suggests signals carrying this type of information in the brain may have a crucial role in shaping neural representations through widespread reciprocal connections in the brain [Salin and Bullier, 1995, Kveraga et al., 2007]. Yet little is known about how these top-down signals orchestrate complex decision making. A central outstanding question in systems neuroscience is how learning is expressed in top-down signals.

Two structures that likely underlie reinforcement learning are the striatum (STR) and prefrontal cortex (PFC). A number of studies and models implicate the striatum in action or choice value representation [Houk, 1995, Suri and Schultz, 1998, Doya, 1910, Nakahara et al., 2001, O'Doherty et al., 2004, Frank et al., 2004, Frank, 2005, Samejima et al., 2005, Pasupathy and Miller, 2005, Histed et al., 2009, Amemori et al., 2011, Daw et al., 2011, Sarvestani et al., 2011, Li and Daw, 2011, Seo et al., 2012, Averbeck and Costa, 2017]. These studies have further suggested that the phasic activity of dopamine, which codes reward prediction errors, drives updates of action value representations stored in the striatum following reward feedback. Several areas in the PFC have also been implicated in dynamic action selection and decision making [Wood and Grafman, 2003, Friston, 2005, Averbeck et al., 2006, Summerfield et al., 2006, Koechlin and Summerfield, 2007, Tsujimoto et al., 2008, Sakai, 2008, Collins and Koechlin, 2012, Botvinick, 2012, Stokes et al., 2013, Lim and Goldman, 2013, Verschure et al., 2014, Botvinick and Weinstein, 2014, Domenech and Koechlin, 2015, Rich and Wallis, 2016, Balaguer et al., 2016, Alexander and Brown, 2018, Aitchison and Lengyel, 2017, Wallis et al., 2019, Radulescu et al., 2019]. These studies further suggest that PFC encodes the memory of past outcomes, plans future actions, and predicts future outcomes. While both striatum and prefrontal cortex have been found to represent action value and choice signals, they differ in degree: value signals were found to be stronger in dSTR than lPFC, while action outcome signals were stronger in lPFC [Pasupathy and Miller, 2005, Samejima et al., 2005, Averbeck et al., 2006, Seo et al., 2012]. In accordance with these findings, we built a joint recurrent network model of the corticostriatal system in which the striatal network represents RL-derived action values and the prefrontal cortex, via recurrent basal ganglia loops, selects appropriate actions based on this signal. We trained this system on a complex decision making task. We also obtained neural recordings from these two regions in two macaques

trained on the same task.

Previous work in the motor system and in prefrontal cortex has shown that insight into the computational mechanisms that underlie complex tasks can be gained by treating the neural population as a dynamical system and studying how its trajectories evolve with time [Rabinovich et al., 2008, Buonomano and Maass, 2009, Sutskever, 2013, Shenoy et al., 2013, Mante et al., 2013, Sussillo and Barak, 2013, Hennequin et al., 2014, Carnevale et al., 2015, Rajan et al., 2016, Gallego et al., 2017, Wang et al., 2017, Chaisangmongkon et al., 2017, Remington et al., 2018, Wang et al., 2018, Yang et al., 2018, Botvinick et al., 2019]. In prefrontal cortex this work has helped shed light on how task execution is driven by dynamics around fixed and slow points neural population space [Mante et al., 2013, Chaisangmongkon et al., 2017]. These studies so far have captured a snapshot of representations when the task is already learned. In the present study, we have used a similar approach to study how representations develop as animals learn to make choices that deliver rewards.

We hypothesized that learning is expressed in latent neural population dynamics. We further hypothesized that a system designed to learn RL-derived action values and select appropriate actions based on them would come to assimilate the function of the corticostriatal system in the brain. If so, the representational structure during task learning should be similar to that found in neural recordings. Moreover, the differing roles assigned to the striatal and prefrontal networks should suffice to induce a difference in representational structure across the two regions in a way that matches the asymmetries in action value and choice representation observed previously [Seo et al., 2012].

We built a recurrent network model of the corticostriatal system composed of two joint networks, a prefrontal network selecting actions and a striatal network representing the value of those actions. We trained this system to autonomously perform a complex sequence-learning task with the same behavioral accuracy as two macaques trained on the same task. Investigating the change in representational structure with learning, we found that movement-sequence representations moved apart from each other in latent space with learning, in both the model and the neural data. We found PFC representations to be more structured and less malleable with learning than dSTR representations. We found that this process was driven by the evolution of gradient landscapes in the networks such that movement-sequence specific fixed point regions moved apart farther from each other in latent space. This increase in distance, or in other words, the increase in the height of the gradient hill between sequence representations makes it increasingly less likely that the wrong action is selected as learning proceeds.

2 Methods and Materials

2.1 Neural Data

The neural data employed here has been previously published in [Seo et al., 2012], though not with the analysis that has been carried out here.

2.1.1 Subjects

Two adult male rhesus monkeys (Macaca mulatta) weighing 5.5–10 kg were used for recordings. All procedures and animal care were conducted in accordance with the Institute of Laboratory Animal Resources Guide for the Care and Use of Laboratory Animals. Experimental procedures for the first animal were in accordance with the United Kingdom Animals (Scientific Procedures) Act of 1986. Procedures for the second animal were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Animal Care and Use Committee of the National Institute of Mental Health (NIMH). Most procedures were equivalent for the two animals except that the UK animal received food pellet rewards while the US animal received juice rewards. The recording chamber (18 mm diameter) was placed over the lateral prefrontal cortex (lPFC) in a sterile surgery using stereotaxic coordinates (AP 26, ML 17 relative to ear-bar zero, in both monkeys) derived from a structural MRI (Fig 1B). This placed the center of the chamber near the caudal tip of the principal sulcus with the FEF in the rear of the chamber.

2.1.2 Task and Stimuli

The two animals performed an oculomotor sequential decision-making task (Fig 1A). A particular trial began when the animals acquired fixation on a green circle (Fixate). If the animal maintained fixation for 500 ms, the green target was replaced by a dynamic pixelating stimulus with a varied proportion of red and blue pixels (the diameter of which was 1 degree of visual angle) and the target stimuli were presented (Stim On). The fixation circle stimulus was generated by randomly choosing the color of each pixel in the stimulus (n = 518 pixels) to be blue (or red) with a probability q. The color of a subset (10%) of the pixels was updated on each video refresh (60 Hz). Whenever a pixel was updated its color was always selected with the same probability q. The set of pixels that was updated was selected randomly on each refresh.

The animal's task was to saccade to the correct target. The animal could make their decision at any time after the target stimuli appeared. After the animal made a saccade to the peripheral target, it had to maintain fixation for 300 ms to signal its decision (first Move + Hold). If the saccade was to the correct target, the target then turned green and the animal had to maintain fixation for an additional 250 ms (Fixate). After this fixation

period, the green target disappeared and two new peripheral targets were presented (Stim On). In the case that the animal made a saccade to the wrong target, the target was extinguished and the animal was forced back to repeat the previous decision step. This was repeated until the animal made the correct choice. For every trial the animal's task was to correctly execute a sequence of three correct decisions for which the animal received either a juice reward (0.1 ml) or a food pellet reward (TestDiet 5TUL 45 mg). After that, a 2000ms inter-trial interval began. The animals always received a reward if it reached the end of the sequence of three correct decisions, even if errors were made along the way. If the animal made a mistake, it only had to repeat the previous decision, it was not forced back to the beginning of the sequence. The full task included both fixed and random conditions, as explained in detail in [Seo et al., 2012]. In the present study, however, only data from the fixed condition was used.

In the fixed condition employed here, the correct spatial sequence of eye movements remained fixed for blocks of eight correct trials 1D). After eight trials were executed without any mistakes, the sequence switched pseudorandomly to a new one. Thus, the animal could draw on its memory to execute a particular sequence, except following a sequence switch.

Every recording session was randomly started with either a fixed or a random set each day and then the two conditions were interleaved. Fixed sets comprised a total of 64 correct trials because the animal had to execute each of the eight sequences correctly 8 times to complete a set. Two runs of the 64 trial set were completed on every recording set so that every one of the eight sequences was repeated twice. comprised two runs of the complete fixed set. The total number of correct and incorrect trials in a particular set depended upon the animal's performance. Neural activity was analyzed if a stable isolation was maintained for a minimum of two sets.

There were eight possible sequences in this task as every trial was composed of three binary decisions 1C). The eight sequences were composed of ten different possible individual movements. Every movement occurred in at least two sequences. We also used several levels of color bias, q as defined above. On most recording days in the fixed sets we used $q \in (0.50, 0.55, 0.60, 0.65)$. The color bias was selected randomly for each movement and was not held constant within a trial. Choices on the 50% color bias condition were rewarded randomly. The sequences were highly overlearned. One animal had 103 total days of training, and the other 92 days before chambers were implanted. The first 5–10 days of this training were devoted to basic fixation and saccade training.

2.1.3 Data Analysis

The neural data analyzed comprised 365 units from dSTR and 479 units from lPFC. We analyzed the activity of individual neurons by fitting an ANOVA model with a 200ms sliding window applied in 25ms steps aligned to

movement onset, as done previously [Seo et al., 2012]. Data was not analyzed if animal failed to maintain fixation or did not saccade to one of the choice targets. We arrived at the number of units reported above by excluding units that did not show a significant effect for the *sequence* factor at any point across the entire recording session as well as units with average firing rates below 1Hz across a recording session. For subsequent analysis, data was pooled across animals and recording sessions and averaged across runs.

To analyze neural population responses, we applied demixed principal component analysis (dPCA) [Brendel et al., 2011, Kobak et al., 2016] to the firing rate traces. As a dimensionality reduction technique, dPCA strives to find a latent representation which captures most of the variance in the data but also expresses the dependence of the representation on different task parameters such as stimuli or decisions. More specifically, it decomposes neural activity into different task parameters, in our case time, sequence, and certainty (and any combination of those) Eq 1. After this decomposition, dPCA finds separate decoder ($\mathbf{D}_\phi$) and encoder ($\mathbf{F}_\phi$) matrices for each task parameter term by minimizing loss function Eq 1 [Brendel et al., 2011, Kobak et al., 2016].

$$\begin{aligned} \mathbf{X} = \mathbf{X}_t + \mathbf{X}_s + \mathbf{X}_c + \mathbf{X}_{sc} + \mathbf{X}_{st} + \mathbf{X}_{ct} &= \sum_\phi \mathbf{X}_\phi + \mathbf{X}_{noise} \\ L_{dPCA} &= \sum_\phi \|\mathbf{X}_\phi - \mathbf{F}_\phi \mathbf{D}_\phi \mathbf{X}\|^2 \end{aligned} \tag{1}$$

The data ($\mathbf{X}$) was smoothed with a Gaussian kernel and then projected into 3-dimensional latent space spanned by the first three vectors of the *sequence*-decoder matrix ($\mathbf{D}_s$).

Distance measures were obtained on the full datasets in full-dimensional neural space, not in the reduced subspace. Euclidean distance between all sequences was computed across all time points for each of the 8 trial repeats and averaged across all possible sequence combinations. In addition, for PFC, distances between sequences were obtained separately for sequences *within* and *across* two clusters (defined by whether a sequence ended in the upper or lower visual hemisphere (sequences S1, S2, S5, S6 and S3, S4, S7, S8, respectively; see Fig.1C). Similarly, Euclidian distance of every sequence to its centroid - defined as the mean across time of a particular sequence trajectory in N-dimensional space with $N = 365$ for dSTR and $N = 479$ for lPFC - was computed across all time points for each of the 8 trial repeats and averaged across sequences.

2.2 Corticostriatal model

2.2.1 Architecture

We jointly trained a connected system of two recurrent neural networks to perform the movement-sequence task (see Methods and Materials). Single-unit dynamics in these networks are governed by: 2.

$$
\begin{aligned}
\tau\,\dot{\mathbf{x}}^s(\mathbf{t}) &= -\mathbf{x}^s(\mathbf{t}) + \mathbf{W}^s_{rr}\mathbf{r}^s(\mathbf{t}) + \mathbf{W}^s_{ir}[\mathbf{u}_a(\mathbf{t}), \mathbf{u}_r(\mathbf{t})] + \boldsymbol{\eta}(\mathbf{t}) \\
\mathbf{r}^s(\mathbf{t}) &= \mathbf{tanh}(\mathbf{x}^s) \\
\tau\,\dot{\mathbf{x}}^p(\mathbf{t}) &= -\mathbf{x}^p(\mathbf{t}) + \mathbf{W}^p_{rr}\mathbf{r}^p(\mathbf{t}) + \mathbf{W}^p_{ir}[\mathbf{r}^s(\mathbf{t}), \mathbf{u}_{ins}(\mathbf{t})] + \boldsymbol{\eta}(\mathbf{t}) \\
\mathbf{r}^p(\mathbf{t}) &= \mathbf{tanh}(\mathbf{x}^p)
\end{aligned}
\tag{2}
$$

where $x^s_i(t)$ and $x^p_i(t)$ are synaptic current variables of unit i at time t in the striatal and prefrontal network, respectively, and activity (firing rate variables r^s_i and r^p_i) is a nonlinear function of x ($r^s_i = \tanh(x^s_i)$ and $r^p_i = \tanh(x^p_i)$), $\mathbf{W}^s_{rr}$ and $\mathbf{W}^p_{rr}$ are the recurrent weight matrices, $\mathbf{W}^s_{ir}$ and $\mathbf{W}^p_{ir}$ are the input weight matrices, and $[\mathbf{u}_a, \mathbf{u}_r]$ and $[\mathbf{r}^s, \mathbf{u}_{ins}]$ are the inputs of the striatal and prefrontal networks, respectively, and added noise $\eta_i(t)$. The striatal and the prefrontal network had $N^s_{rr} = 1300$ and $N^p_{rr} = 1000$ units, respectively. The connectivity weight matrices $\mathbf{W}^s_{rr}$ and $\mathbf{W}^p_{rr}$ were initially drawn from the standard normal distribution and multiplied by a scaling factor of $\sqrt{N^s_{rr}}$ with $N_{rr} = 1300$ and $\sqrt{N^p_{rr}}$ with $N^p_{rr} = 1000$. The neural time constant was $\tau = 10ms$. Each unit receives an independent white noise input, η_i, with zero mean and $SD = 0.01$. Inputs are fed into the networks through the input weight matrics $\mathbf{W}^s_{ir}$ and $\mathbf{W}^p_{ir}$ which were initially drawn from the standard normal distribution and multiplied by a scaling factor of $\sqrt{N^s_{ir}}$ with $N^s_{ir} = 20$ and $\sqrt{N^p_{ir}}$ with $N^p_{ir} = 510$ for the prefrontal network. Otherwise, the model parameters were set in the same range as [Mante et al., 2013].

The striatal network receives a 15-dimensional input vector, $\mathbf{u}_s = [\mathbf{u}_a, \mathbf{u}_r]$, composed of a 10-dimensional vector $\mathbf{u}_a$ specifying actions taken and a 5-dimensional vector $\mathbf{u}_r$ specifying rewards received for those actions. The prefrontal network receives a 510-dimensional input, $\mathbf{u}_p = [\mathbf{r}^s, \mathbf{u}_{ins}]$, composed of a 500-dimensional vector $\mathbf{r}^s$ of activations from a subste of the hidden units in the striatal network, together with a 10-dimensional instruction vector $\mathbf{u}_{ins}$ specifying the interval for the fixation period (Fixate) and when to move towards or hold the target (Move + Hold).

Outputs for each of the networks were linearly read out from the synaptic currents of the recurrent circuit (Eq 3).

$$
\begin{aligned}
\mathbf{y}^s &= \mathbf{W}^s_{ro}\mathbf{x}^s \\
\mathbf{y}^p &= \mathbf{W}^p_{ro}\mathbf{x}^p
\end{aligned}
\tag{3}
$$

The striatal network outputs a 10-dimensional output vector, $\mathbf{y}^s$, of action values derived from TD learning (see below). The prefrontal network outputs a 11-dimensional output vector, $\mathbf{y}_p = [\mathbf{y}_a, y_v]$ composed of a 10-dimensional vector $\mathbf{y}_a$ of actions and an additional unit coding for the visual hemisphere (upper or lower) to

which the last movement of a particular sequence belonged.

2.2.2 Training & Coding

All synaptic weight matrices ($\mathbf{W}_{rr}$, $\mathbf{W}_{ir}$, and $\mathbf{W}_{ro}$) are updated with the gradient of loss function (Eq 4), which is designed to minimize the square of the difference between network and target output:

$$\mathbf{l} = 2\sum_{k=1}^{K}\sum_{t=0}^{T}\sum_{i=1}^{N_{rrs}}(\hat{y}_i - y_i)^2 + \sum_{k=1}^{K}\sum_{t=0}^{T}\sum_{i=1}^{N_{rrp}}(\hat{y}_i - y_i)^2 \tag{4}$$

The error is thus obtained by obtaining the difference between network output y and target output $\hat{y}$ and summing over all trials K in a batch (with $K = 10$), time points T and recurrent units N_{rr} (with $N_{rrs} = 1300$ and $N_{rrp} = 1000$ for the striatal and prefrontal network, respectively). The total loss function l (Eq 4) was obtained by combining the loss terms from the striatal and the prefrontal network while assigning double weight to the striatal loss term. The striatal and the prefrontal networks were jointly trained by obtaining the gradient of the combined loss function (Eq 4) through automatic differentiation with *autograd* [Maclaurin et al.,] and custom implementations of accelerated functions (i.e. Adam optimization) with GPU-based computations using *JAX* [Johnson et al., 2018]. The network was trained with an initial learning rate of $\alpha = 0.001$ for 10 steps with 1000 iterations each, while the learning rate was decayed by $\frac{2}{3}$ at every step. After this training phase, all the synaptic weight matrices were fixed.

Actions ($\mathbf{u}_a/\mathbf{y}_p$) and action values ($\mathbf{y}_s$) were coded as 10-D vectors in which every unit coded for one of the movement choice options (see Methods and Materials - Task and Stimuli). So, for instance, for the first binary choice option (Fig 1C, S1-center) one unit coded the right movement choice and the other the left movement choice. Rewards ($\mathbf{u}_r$) and instructions ($\mathbf{u}_{ins}$) were coded as 5-D vectors with every unit coding for reward delivered at one of the 5 decision points (center, upwards, downwards, upper, lower; Fig 1C). Actions and rewards were coded as pulses, with reward pulses lasting $\frac{2}{10}$th of the action pulse interval and being delivered after the end of an action pulse.

Rewards drove the update of action values according to a temporal-difference reinforcement learning paradigm (TD learning) (Eq 5) [Sutton and Barto, 2018].

$$\begin{aligned} \mathbf{V}(\mathbf{S}_t) \longleftarrow \beta\mathbf{V}(\mathbf{S}_t) + \alpha\left[\mathbf{R}_{t+1} + \gamma\mathbf{V}(\mathbf{S}_{t+1}) - \mathbf{V}(\mathbf{S}_t)\right] \\ \mathbf{p(t)}_{choice} = softmax(\rho\mathbf{V}(\mathbf{S}_t)) \end{aligned} \tag{5}$$

The learning rate parameter α, the discount factor γ and an additional inverse temperature parameter ρ

were fitted to one of the training sessions of monkey 1 using *fminsearch* in Matlab. The decay parameter β was set to 0.8. The values obtained for the parameters were $\alpha = 0.9864$, $\gamma = 0.6120$, and $\rho = 7.328$.

The training data set for the corticostriatal model system was mainly composed of real behavioral data from across all training sessions of the two monkeys. Action values ($\mathbf{y}_s$) were derived by feeding the actions and rewards the animals received through the TD-learning algorithm (Eq 5). A subset of 25 blocks from the real data was left out as a test set. Additionally, the training data set was augmented with artificial data generated by randomly assigning a particular sequence to the current block and drawing action movement outcomes according to the same error probabilities that the animals displayed in the task (see Fig 4C-Behavior). During training, a batch composed of 10 blocks was randomly chosen from across the entire training set at each step.

2.2.3 Autonomously taking actions

After training, we also tested the model's ability to autonomously produce movement-sequence blocks. We did this by obtaining trial-by-trial value and action estimates and feeding the outputted action back into the system at the next step. We obtained initial value and action estimates from the striatal and prefrontal network, respectively, by setting the initial striatal input vector ($\mathbf{u}_s$) to zero (plus white noise, $\boldsymbol{\eta}$). In order to decode actions, we chose the action corresponding to the highest valued movement direction ($a* = max(y_p)$).

If the correct action was outputted, the prefrontal output action vector ($\mathbf{y}_a$) was fed back into the striatal network as the action input vector ($\mathbf{u}_a$) together with the corresponding reward vector ($\mathbf{u}_r$). If the wrong action was taken, the action was recorded and the output action vector was set to the correct target vector before feeding it back into the striatal network at the next step. This was done to ensure the network produced the correct movement before proceeding to the next step, analogous to the animals in the task who were forced to repeat wrong movements until correct before proceeding. The last output action vector at the end of a particular block was assigned as the first input action vector ($\mathbf{u}_a$) to the striatal network in the new block.

2.2.4 Model Analysis

All the reported model-related findings are derived from the test set which had not been exposed to the system during training. The neural population activity from the striatal and prefrontal networks of the corticostriatal model was imaged in the same way as the real neural recordings, by projecting activity into a 3-dimensional latent space spanned by the first three vectors of the *sequence*-decoding matrix ($\mathbf{D}_s$) obtained through dPCA (Eq 1) [Brendel et al., 2011, Kobak et al., 2016].

In order to image the gradient vector field around the latent trajectories, we first projected neural population activity into latent space and obtained a 3-dimensional mesh of points ($\mathbf{X}^*$) around the sample trajectories.

Then we projected this mesh of points back out into neural population space using the *stimulus*-encoder matrix ($\mathbf{F}_s$) (Eq 1). We started off the networks at each of these points by providing them as the initial vector of firing rates ($\mathbf{r(t)}$, Eq 2), and iterated the network through a whole block of movement-sequence trials. For illustration purposes, we picked particular points in time across the block and imaged the magnitude of the gradient from one step to the next for each of the points in the mesh. In order to ensure that the trough of the gradient manifold really pointed to fixed point regions for those particular moments in time, we kept the input fixed and continued iterating the network for up to 10 iterations (which corresponded to the length of the fixation period) to make sure the location of the trough remained fixed on the timescale of the network. If the location where the gradient vector field magnitude was at its minimum remained fixed during this iteration period, we labelled that location as a fixed point. In order to image the gradient vector field of two particular sequences together, we obtained the joint gradient manifold by taking the point-wise minimum across the two sequences' manifolds.

In order to calculate the distance between the minima of the gradient vector fields of two different sequences, we first confirmed the location of the fixed points in 3-dimensional latent space as described above. Then, we used Dijkstra's algorithm to obtain the minimal path length along the joint gradient manifold between two particular sequences' fixed points. We repeated this for all possible sequence combinations in the test set and averaged the result. In order to calculate the distance between sequences in latent space as well as the distance of a particular sequence to its centroid, we first projected the data into 10-dimensional dPC-latent space (for computational reasons) and then proceeded the same way as in the neural data (see Methods and Materials - Neural Data - Data Analysis). We found no difference in these metrics when projecting into 10- or 20-dimensional latent space.

3 Results

We investigated dynamics during task learning in the corticostriatal system (Fig.1). The task consisted of a sequence of three movements which the animal had to execute by saccading to the correct target (Fig.1A). There were a total of 8 different possible arrangements (or sequences) of 3 sequential movements (Fig.1C). These sequences, in turn, were arranged into blocks in which one particular sequence would remain fixed for the duration of a block. Thus the animal was able to use his memory of which sequence had been correct in the previous trial to improve his performance across a particular block. At the start of a new block, a new sequence was chosen randomly (Fig.1D, see Methods and Materials).

Two animals performed this task while we obtained neural recordings from lateral prefrontal cortex (lPFC) and dorsal striatum (dSTR) (Fig.1B, see Methods and Materials) [Seo et al., 2012]. To investigate learning dynamics, we built a model of the corticostriatal system. In this model, the prefrontal network was trained to

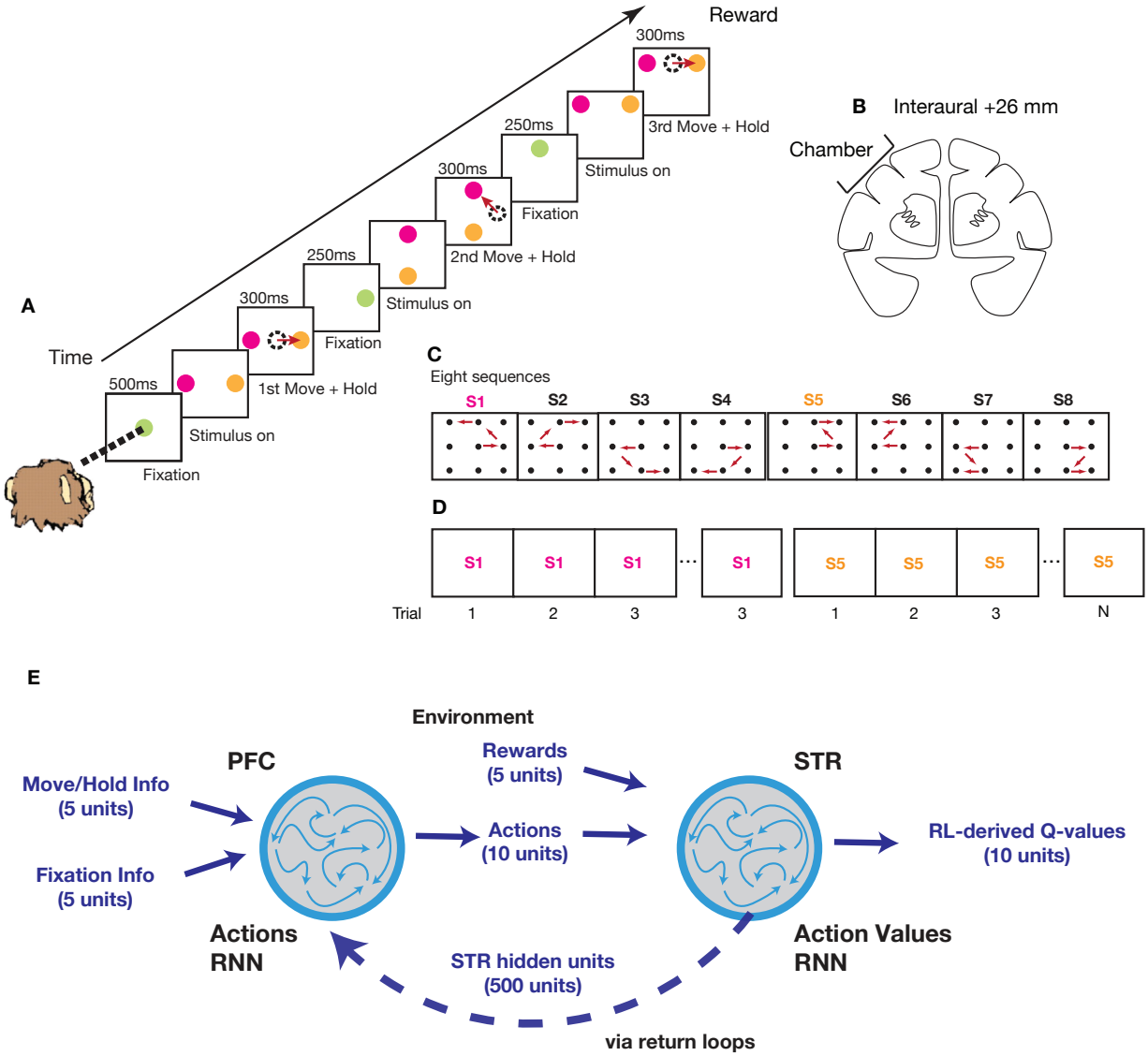

Fig 1. Overview

produce actions, while the striatal network mapped actions and rewards onto values (Fig.1E). The prefrontal network, in turn, received inputs from a subset of units in the striatal network so that action value representations were used to drive action selection. The prefrontal network received additional instruction inputs signifying fixation and move/hold periods. The two networks in this system were jointly trained using actions and rewards from the two animals' real recorded behavior. Corresponding action values were generated by feeding actions and rewards through a temporal difference reinforcement learning algorithm (TD learning). The network setup and training procedure are described in detail in Methods and Materials.

After training, the corticostriatal model system learned to produce correct movement sequences (Fig.2). In this particular example, sequence 5 (consisting of rightwards, upwards, and rightwards movements) was the correct sequence for the current block. The prefrontal network units coding for these particular movements correctly output an action pulse (light blue). Since this was the correct action, the network receives a reward input (dark blue). The reward, in turn, causes the value signal in the striatal network (red) to increase for these particular, rewarded movements. Meanwhile, the output of the prefrontal and striatal network units coding for other movement directions remains flat. Altogether, the system learned to produce movement sequences with the correct action and action value output.

Stringing together several trials of movement sequences, one can obtain the two networks' output for a whole block of movement sequences (Fig.3). There are a total of 10 output units, one for each of the available movement directions (see Methods and Materials). The striatal network received action and reward inputs from two blocks which were part of the test set (and had not been presented to the network during training). Action outputs (by the prefrontal network, in light blue), action value (or Q-value) outputs (by the striatal network, in red) and rewards (received from the environment, in dark blue) are all imaged together for these two sequential blocks. The correct sequence differed from one block to the next (sequence S5 for the first block, and S8 for the second).

Thus, during the first block (trials T6*, T7* and T8*), the three output units on the top right are active, representing a movement from the center to the right, followed by a movement from the right upwards, and a movement to the right at the top. The prefrontal network signals the correct movement sequence (S5) through sequential action pulses (light blue) in these three output units. Rewards are presented as short pulses at the end of an action (dark blue). The striatal network subsequently outputs the corresponding action value signal (red). The value signal decays over time, but recovers when correctly executed trials follow upon each other (as is the case with trials T6*, T7* and T8*, Fig.3).

As the new block begins (dotted line), the correct sequence (S8) is now signalled by the three output units on the bottom right, representing a movement from the center to the right, followed by a movement from the right downwards, and a movement to the right at the bottom. Meanwhile, the striatal network's value signal in

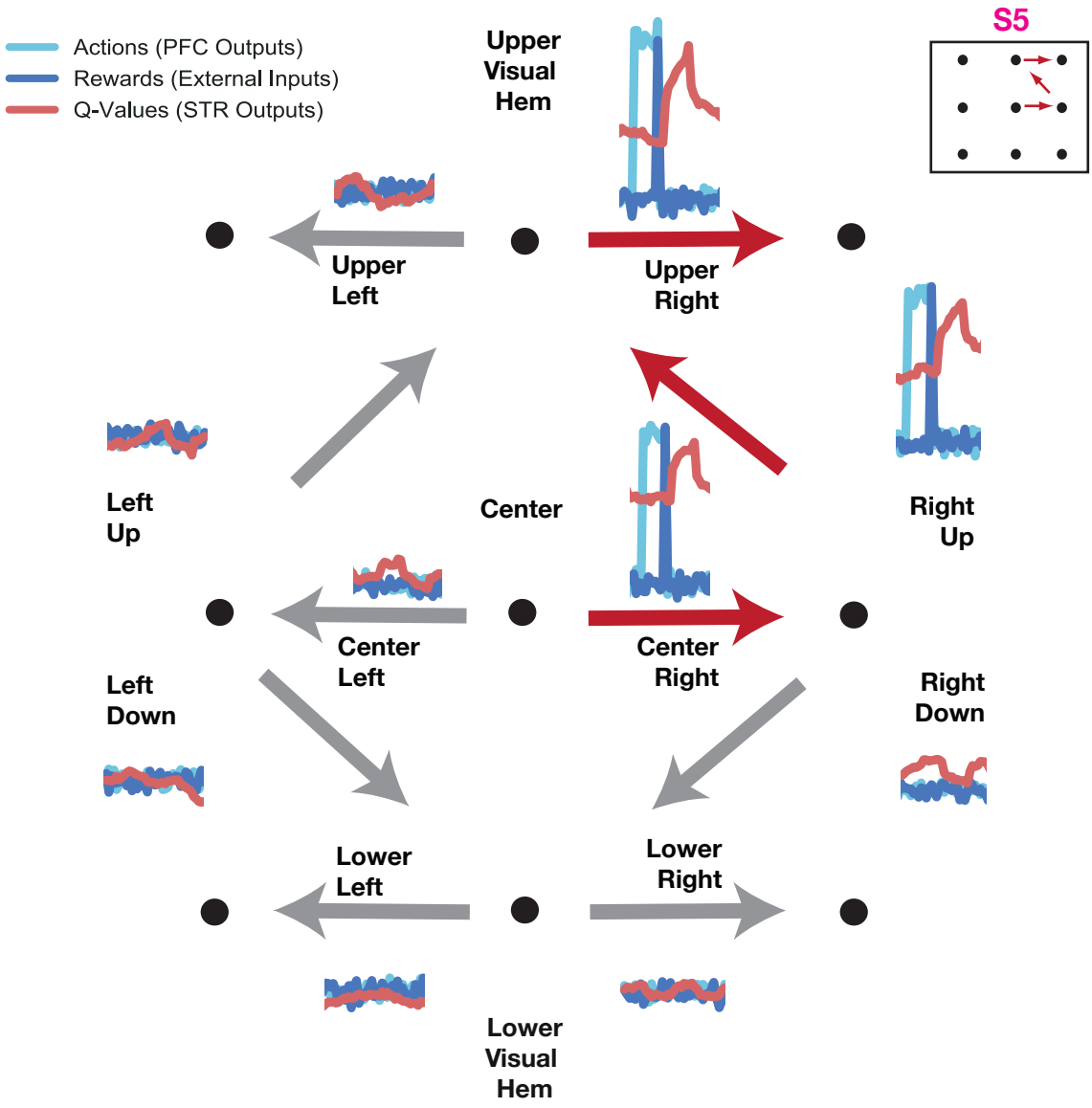
Actions (PFC Outputs)
Rewards (External Inputs)
Q-Values (STR Outputs)
S5
Upper Visual Hem
Upper Left
Upper Right
Left Up
Center
Right Up
Center Left
Center Right
Left Down
Right Down
Lower Left
Lower Right
Lower Visual Hem

Fig 2

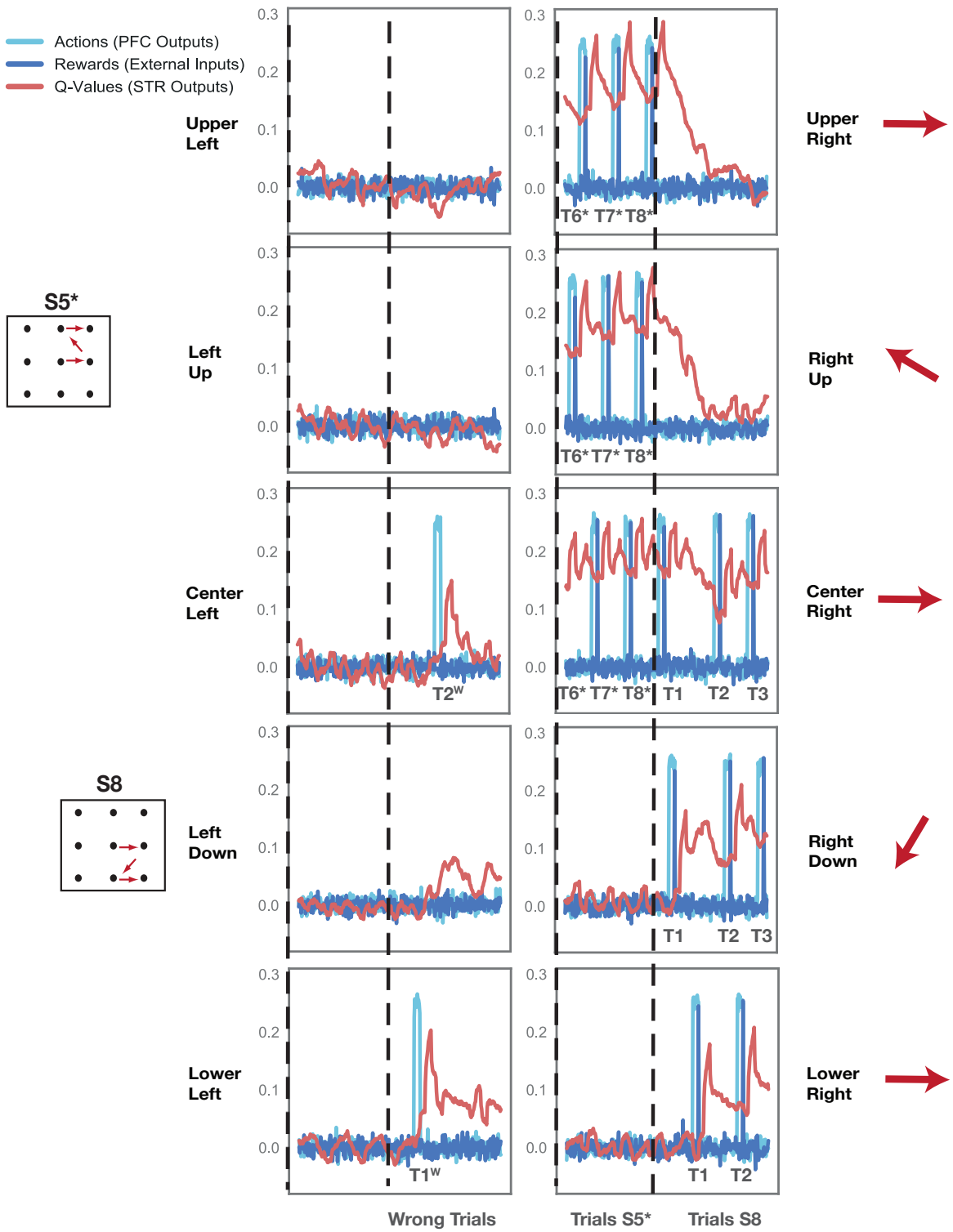
Actions (PFC Outputs)
Rewards (External Inputs)
Q-Values (STR Outputs)
S5*
S8
Upper Left
Left Up
Center Left
Left Down
Lower Left
Upper Right
Right Up
Center Right
Right Down
Lower Right
0.3
0.2
0.1
0.0
T6* T7* T8*
T2W
T1
T2
T3
T1W
Wrong Trials
Trials S5*
Trials S8

Fig 3

the units which are no longer correct (right up and upper right) decays back towards zero.

At trials T1 and T2 of the new block, wrong actions were taken (for the last movement of trial 1 in the lower left, $T1^w$, and for the first movement of trial 2 in the center left, $T2^w$). This subsequently caused the striatal value signal to erroneously increase for those wrong actions. The particular action step for which the wrong movement occurred was repeated immediately after (just as it was for the animals in the actual task, see Methods and Materials), and the correct action was produced in the second attempt (T1 lower right, and T2 center right). The striatal value signal subsequently increases for the correct movement direction (lower right and center right), while the value signals for the wrong movement directions decay (lower left and center left). Altogether, the system of networks learned to produce movement-sequence blocks with a distribution of correct and erroneous actions that approximates the behavior of the animals during the real task.

To analyze the system's performance in more detail, we imaged striatal and prefrontal outputs together with their target outputs for a sample block transition (Fig.4A,B). Output traces (red) generally follow target traces (green). To quantify this, we computed the mean squared error between targets and outputs over the duration of a block averaged over all blocks in the test set (Fig.4F,G). The largest difference between outputs and targets occurs for the first trial, when the correct sequence is unknown, and the difference becomes smaller over the course of the block for both prefrontal (Fig.4F) and striatal networks (Fig.4G).

We also imaged the eigenvalue spectra of the two networks' weight matrices after training (Fig.4D,E). We found that variance is spread across many dimensions in the prefrontal network (Fig.4D), while most of the variance is concentrated in around 5 large eigenvalues in the striatal network (Fig.4E).

We also determined the behavior of the model system by measuring the fraction of correct decisions over the course of the movement-sequence block (Fig.4C). The recorded behavioral data from the animals (solid line, [Seo et al., 2012]) shows chance performance at the beginning of the block, and rapid, steady improvement over the course of the block. In order to determine the fraction of correct decisions of the model system, we let the networks produce movement-sequences autonomously (see Methods and Materials-Corticostriatal model). That is, we used the action outputs of the prefrontal network during the current trial as inputs to the striatal network in the next trial. In this way we obtained a distribution of activations for the next movement step in the prefrontal output units, from which we decoded the predicted action. When averaged over 25 blocks, the fraction of correct decisions of the autonomous model system (dotted line) approximates that of the animals' behavior (solid line). We also plotted the fraction of correct decisions obtained from the fit of the TD-learning algorithm (dashed line), and it agrees well with real behavior. Altogether, this shows the model system was able to capture the animals' behavior in this task.

In order to study how neural representations evolve with learning, we imaged PFC neural population

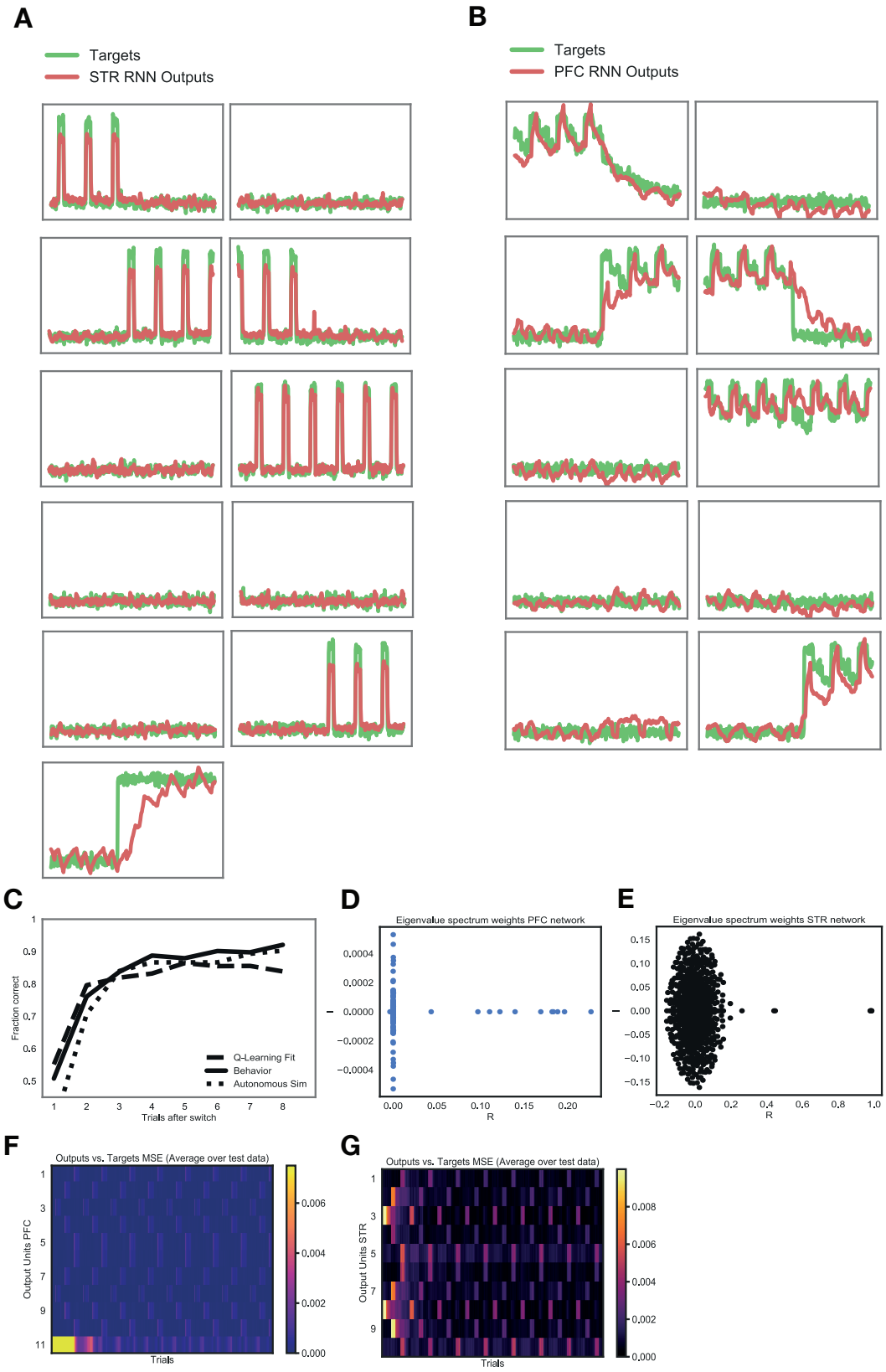
A
Targets
STR RNN Outputs
B
Targets
PFC RNN Outputs
C
Fraction correct
Trials after switch
Q-Learning Fit
Behavior
Autonomous Sim
D
Eigenvalue spectrum weights PFC network
R
E
Eigenvalue spectrum weights STR network
R
F
Outputs vs. Targets MSE (Average over test data)
Output Units PFC
Trials
G
Outputs vs. Targets MSE (Average over test data)
Output Units STR
Trials

Fig 4

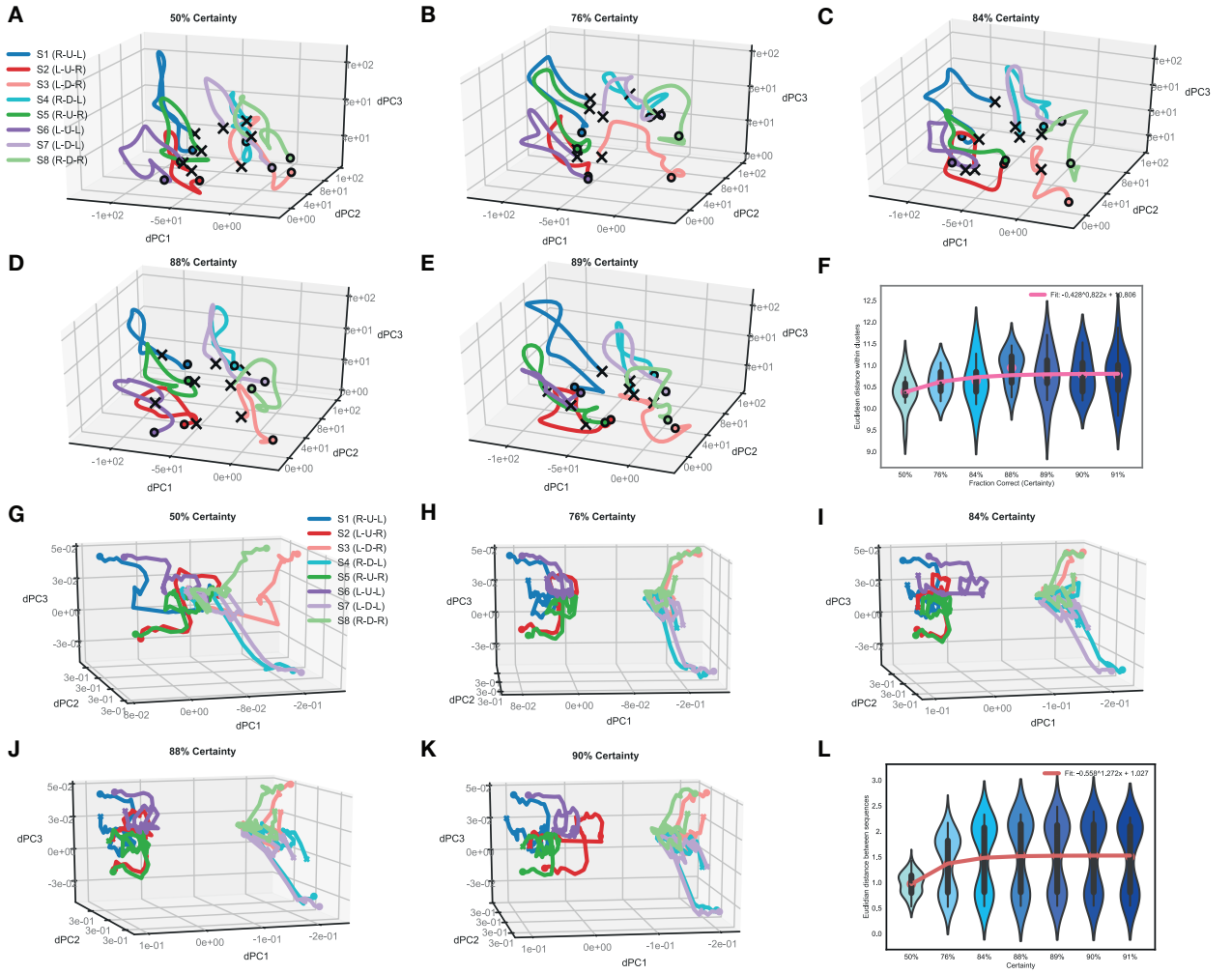

Fig 5

activity in 3-dimensional latent space using demixed principal component analysis (see Methods and Materials). Trajectories from neural recordings in lPFC (Fig.5A-F) are plotted alongside trajectories from the prefrontal network of the corticostriatal model (Fig.5G-L). The plots capture representations for different trials over the course of a block (Fig.1D), as the animals progress from 50% certainty (or fraction correct) at the start of the block to close to 90% certainty by the fifth trial into the block (Fig.5C).

Sequence representations from neural recordings in lPFC (Fig.5A-F) showed a separation by visual hemisphere: sequences S1, S2, S5 and S6 which progressed along the upper visual hemisphere (Fig.1C) were clustered to the left while the remaining sequences which progressed along the lower visual hemisphere were clustered to the right. This separation appeared to be present from the very start of a block and was maintained with increasing certainty (Fig.5A-E). Meanwhile, sequence trajectories within each particular cluster separated

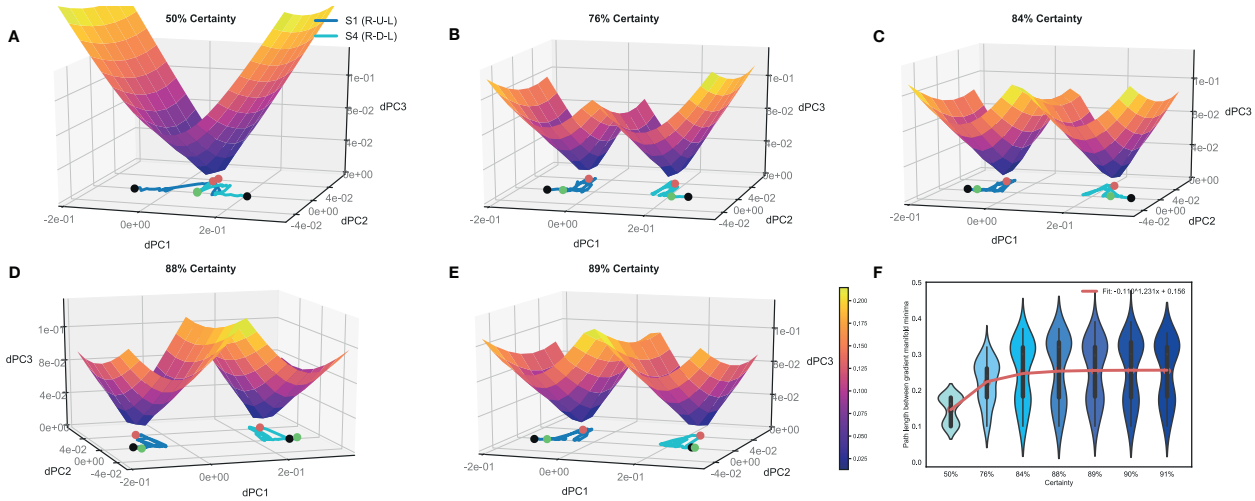

Fig 6

more from each other with increasing certainty (i.e. Fig.5A vs. Fig.5B). To capture this effect, we computed the Euclidean distance between sequences within clusters in neural population space (in the full-dimensional space of recordings, not in the reduced latent space; see Methods and Materials). We found this measure increased with certainty as learning progressed over the block (Fig.5F). Sequence representations from the prefrontal network model (Fig.5G-L) showed the same clustering by visual hemisphere. In contrast to the real data, however, model trajectories showed less separation at the start during the 1st trial (Fig.5G vs Fig.5A). Overall, though, we also found Euclidian distance between sequences within clusters to be increasing for the model trajectories (Fig.5L).

To study what underlies increasing separation of movement-sequence representations with learning, we probed the prefrontal model network further (Fig.6). We obtained the gradient vector field around the latent sequence trajectories as learning progressed across the block (see Methods and Materials - Model Analysis). We obtained the common vector field across two different movement-sequence trajectories in latent space by taking the point-wise minimum across the gradient vector fields of the individual sequences, and imaged this common gradient manifold for increasing levels of certainty across a block (Fig.6A-E). The gradient vector field shows how activity evolves at different points in neural latent space in the vicinity of sequence trajectories. Neural activity has a high propensity to be pushed away from locations where the magnitude of the gradient is high (yellow), and remain in locations where the magnitude is low (dark blue). We ascertained that the troughs of the gradient manifold point to fixed points in neural activity (see Methods and Materials - Model Analysis).

Observing the evolution of common gradient manifold with learning (Fig.6 A-E) for a particular point along the movement trajectory (red point), one notices how fixed point regions for different regions become increasingly

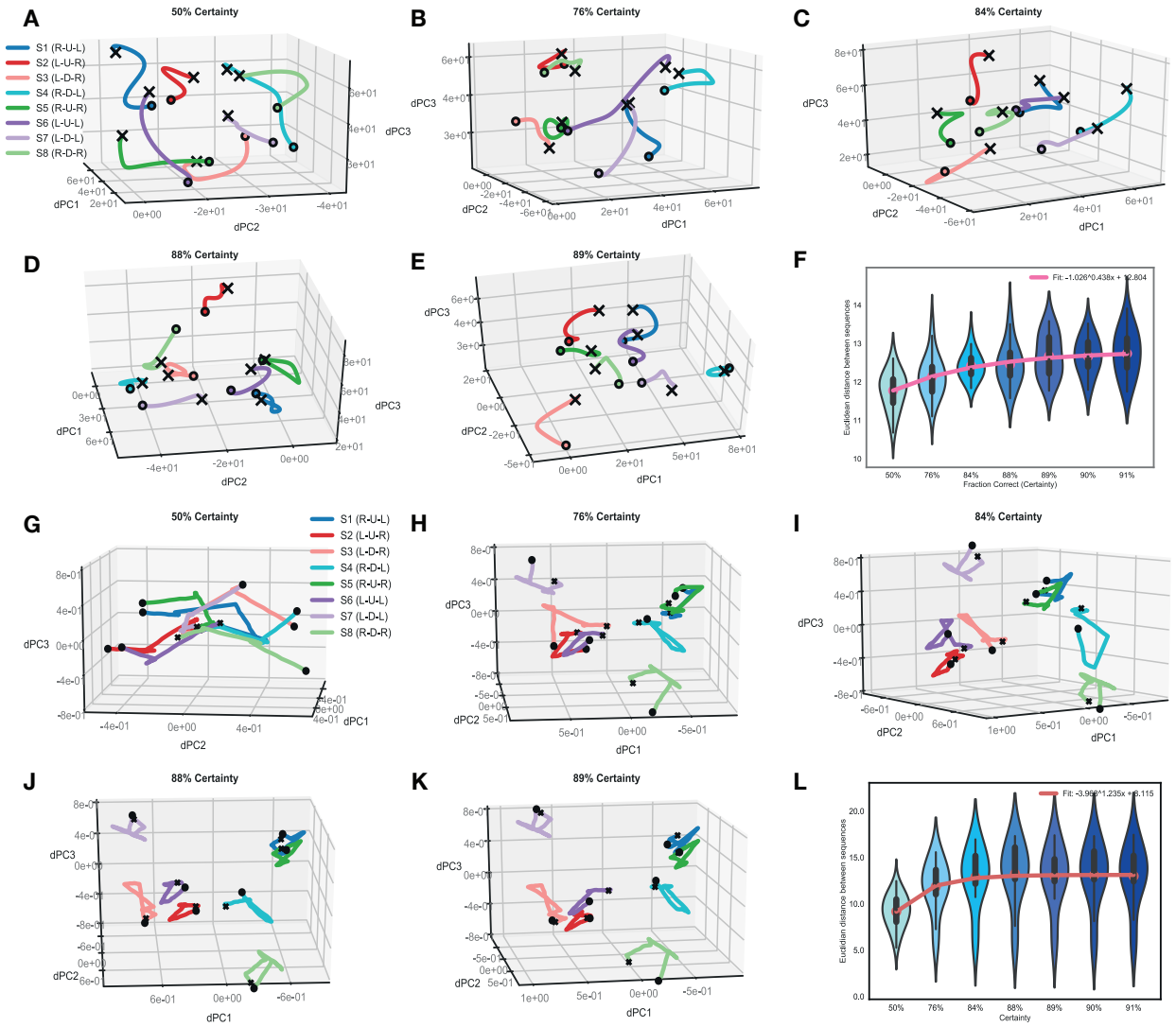
A
50% Certainty
S1 (R-U-L)
S2 (L-U-R)
S3 (L-D-R)
S4 (R-D-L)
S5 (R-U-R)
S6 (L-U-L)
S7 (L-D-L)
S8 (R-D-R)
dPC1
dPC2
dPC3
B
76% Certainty
C
84% Certainty
D
88% Certainty
E
89% Certainty
F
Fit: -1.026^0.438x + 12.804
Euclidean distance between sequences
Fraction Correct (Certainty)
50%
76%
84%
88%
89%
90%
91%
G
50% Certainty
H
76% Certainty
I
84% Certainty
J
88% Certainty
K
89% Certainty
L
Euclidian distance between sequences
Certainty

Fig 7

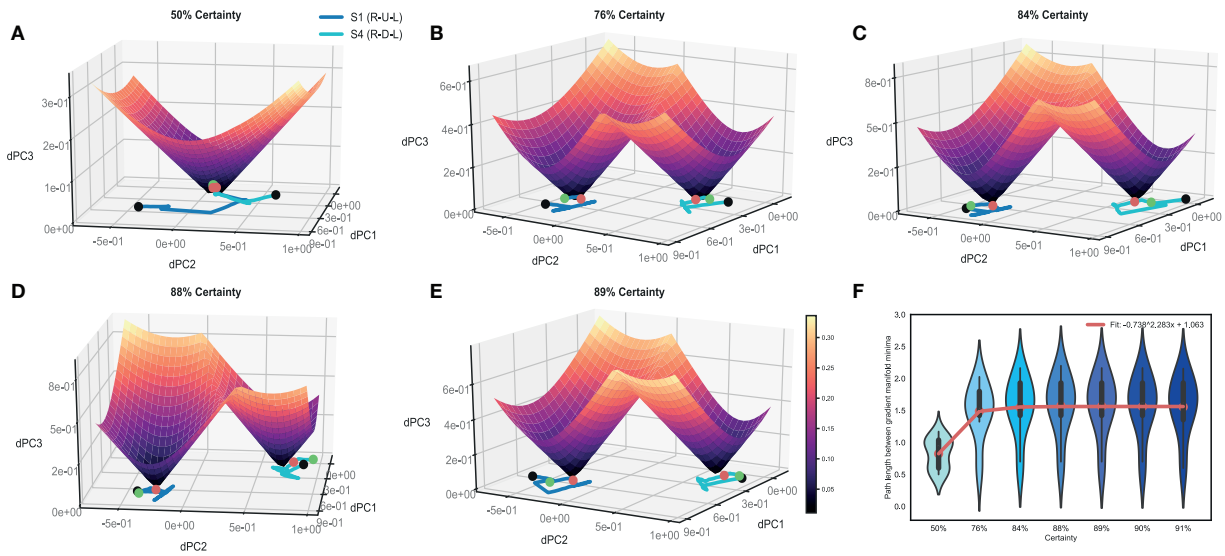

Fig 8

well separated. Along with this separation, the ridge in the gradient manifold between the two sequences' fixed point region heightens with increasing certainty during learning. As this happens, it becomes increasingly less likely to commit errors: the gradient along a particular side of the ridge drives activity more strongly towards a particular sequence's fixed point region, so the chance to end up in another's decreases with learning. To quantify this effect, we determined the minimal path length between the various sequence's fixed point regions as learning progressed (Fig.6F; see Methods and Materials - Model Analysis). We observed an increase in minimal path length between fixed point regions with increasing certainty during learning (Fig.6 F). Altogether, we established that the gradient in the network pushes apart fixed point regions for different movement-sequences as learning progresses, underlying increasing behavioral accuracy.

We also examined how neural representations evolve with learning in the striatum (Fig.7). Trajectories from neural recordings in dSTR (Fig.7A-F) are plotted alongside trajectories from the striatal network of the corticostriatal model (Fig.7G-L). Sequence representations in the dSTR (Fig.7A-E) do not display any particular clustering by visual hemisphere, unlike representations in lPFC (Fig.7A-E). Sequence representations are instead scattered around in latent space. As learning progresses, sequence representations spread further apart from each other (i.e. along dPC2 from 5o% certainty, Fig.7A, to 76% certainty, Fig.7B). We computed the Euclidean distance between all sequences in neural population space and found it to be increasing with learning (Fig.7F). Trajectories from the striatal network of the corticostriatal model (Fig.7G-L) were also scattered around latent spice, like in the neural recordings (Fig.7A-E). Model representations at the start of the block (Fig.7G), however,

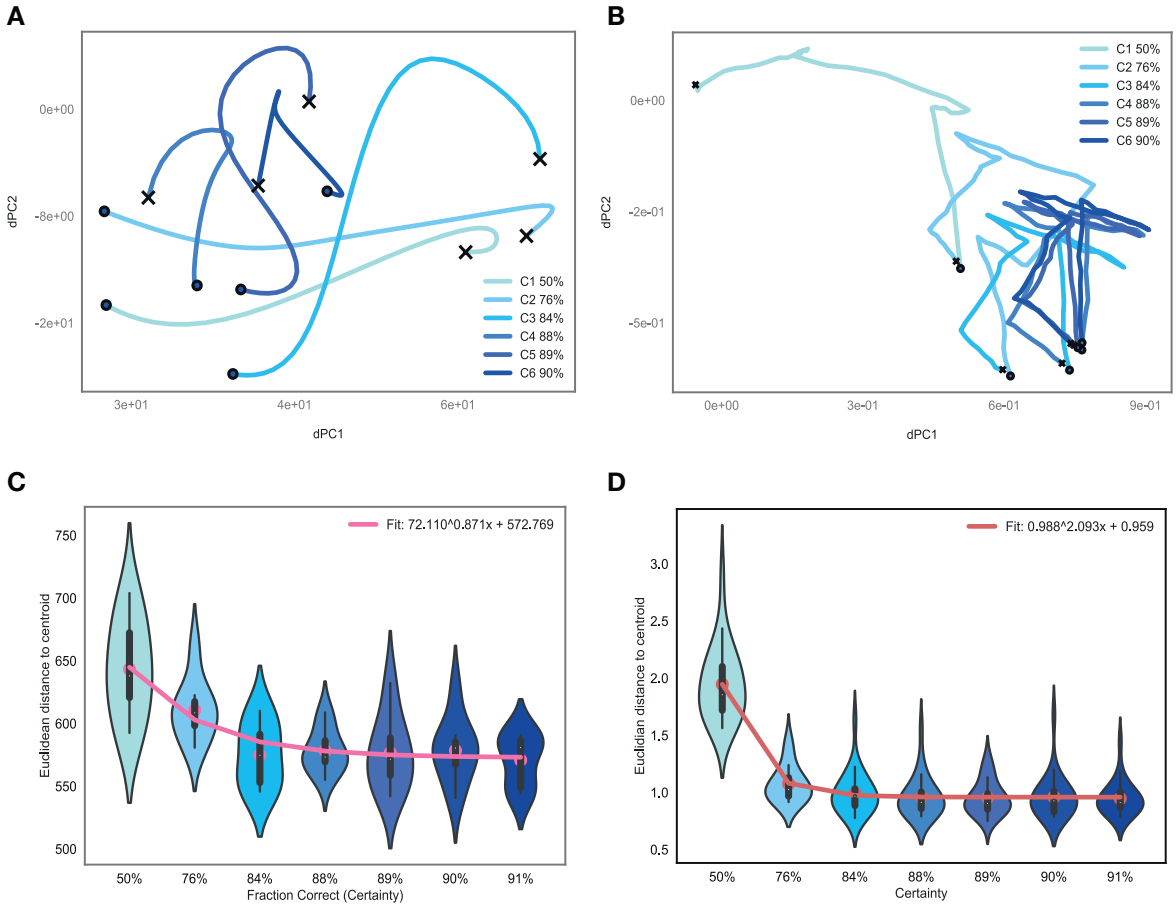

Fig 9

were closer together than neural representations (Fig.7A). Like in the recordings, sequence representations in the model spread further apart from each other as learning progressed. To quantify this effect, we again computed the Euclidean distance between sequences (see Methods and Materials) and found it to be increasing with learning (Fig.7L), as in the neural data (Fig.7F).

We also examined how the shape of a particular movement-sequence representation in latent space changes with learning in the STR (Fig.9). We observed that trajectories become more compact with learning (from lighter to darker blue), in neural recordings from dSTR (Fig.9A) and in the striatal model network (Fig.9B). To quantify this effect, we computed the Euclidean distance of a particular sequence to its centroid for increasing certainty levels (see Methods and Materials) and found this measure to be decreasing with learning in both neural recordings (Fig.9C) and in the striatal model (Fig.9D). Altogether, we found sequence representations became more compact with learning in STR recordings and in the model.

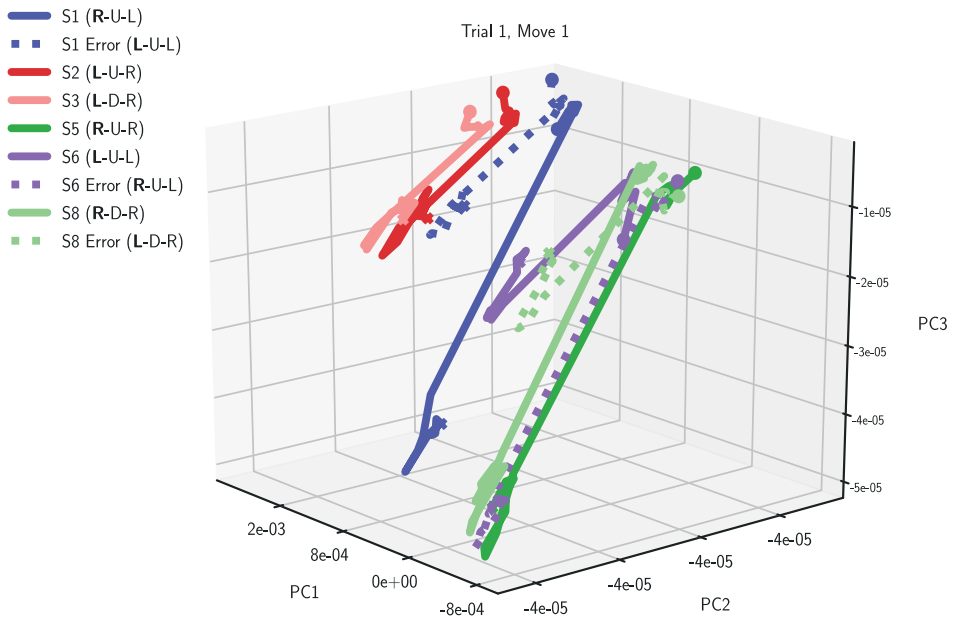

Fig 10

We further compared latent representations for correct and wrong movements in the prefrontal model network (Fig.10). We imaged representations of the first movement during the first trial in the block (when error rate is highest) in 3-dimensional dPC-latent space, for a few different sample sequences. We found that when the wrong movement is executed (i.e. S1 Error, dotted blue line), the representation moves away from that of the correct movement (right move, R in bold, for sequence S1, solid blue line) and closer to that of sequences which share the same executed movement (left move, L in bold, for sequence S2, solid red line, and sequence S3, solid rose line). Similarly, the movement representation for the wrong move in sequence 8 (S8, dotted light green line) moves away from the correct move trajectory (solid light green line) and closer to the movement representation of sequence 6 (solid magenta line) which shares the same movement direction.

4 Discussion

The results contribute four important insights into the function of the corticostriatal system in the brain. First, our model revealed that task learning in neural populations proceeds by shaping the gradient landscape such that fixed point regions corresponding to different task dimensions are pushed farther apart from each other. This accompanied increasing behavioral accuracy in task performance. Second, representations in our corticostriatal

model system in which a striatal network encoded action values and a prefrontal network selected actions approximated representations recorded from the dSTR and lPFC of the macaque brain. This suggests the function of the corticostriatal system in the brain is well approximated by striatal encoding of RL-derived action values and prefrontal encoding of selected actions. In both, recordings and model system, we found PFC representations to be clustered by visual hemisphere and more fixed with learning than STR representations, which displayed no particular organizational structure. Third, the corticostriatal model system was able to autonomously perform the task with the same behavioral accuracy as the animals, suggesting this system is sufficient to perform a complex learning task while approximating neural representations well. Fourth, we found that STR sequence representations spread apart from each other while also becoming more compact with learning. Moreover, we found that when wrong movements occurred, sequence representations moved closer to other sequence representations which shared the executed movement. Altogether, our findings offer insight into the learning process in neural populations and provide testable predictions for further research.

The finding that lPFC representations are less malleable with learning and more structured than dSTR representations agrees with previous results [Pasupathy and Miller, 2005, Samejima et al., 2005, Seo et al., 2012]. Seo et. al. (2012) found less units in lPFC that represented a reinforcement learning (RL) variable and more that encoded sequence information. Conversely, more units in dSTR showed a significant effect for RL and less units showed an effect for sequence information.

At a higher level, the corticostriatal model system respects neuroscientific evidence that implicates the striatum in action value representation [Pasupathy and Miller, 2005, Samejima et al., 2005, Averbeck and Costa, 2017] and the PFC in action selection [Averbeck et al., 2006]. At a lower level, the methods used to train this system are not biologically validated, similar to previous approaches [Mante et al., 2013, Yamins et al., 2014, Chaisangmongkon et al., 2017, Yang et al., 2018]. To make it more plausible, one could implement this system with spiking networks [Nicola and Clopath, 2017]. Another approach may be to use a reinforcement learning paradigm, rather than the gradient of an error signal, to train the network system [Song et al., 2017]. However, the latent dynamics underlying task learning uncovered here evolve over a longer timescale than the learning dynamics underlying network training, so findings are likely not impacted by different training protocols.

We found that latent representations in STR also become more compact at the same time as sequence-specific trajectories spread further apart from each other in latent space. This effect could be partially driven by changes in the mean and variance of the neural population firing rate with learning, as well as by changes in higher order statistics. Previously it was found that the Fano factor, a measure of variability (variance of spike count divided by its mean), decreases in prefrontal neurons with learning [Qi and Constantinidis, 2012]. Also, changes in firing patterns within the neural population with learning - such as changes in synchronous

firing [Baeg et al., 2007] - might be responsible for the changes in latent representation observed here.

5 Acknowledgments

This work was supported by a Wellcome Trust NIH-fellowship (C.D.M. & S.S.) and by NIH grant ZIA MH002928-01 (B.B.A.).

6 Competing Interests

The authors report no conflicts of interest.

References

Aitchison and Lengyel, 2017. Aitchison, L. and Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227.

Alexander and Brown, 2018. Alexander, W. H. and Brown, J. W. (2018). Frontal cortex function as derived from hierarchical predictive coding. *Scientific Reports*, pages 1–11.

Amemori et al., 2011. Amemori, K.-i., Graybiel, A. M., and Gibb, L. G. (2011). Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments. *Frontiers in Human Neuroscience*, 5(47):1–20.

Averbeck and Costa, 2017. Averbeck, B. B. and Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*, 20(4):505–512.

Averbeck et al., 2006. Averbeck, B. B., Sohn, J.-W., and Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, 9(2):276–282.

Baeg et al., 2007. Baeg, E. H., Kim, Y. B., Kim, J., Ghim, J. W., Kim, J. J., and Jung, M. W. (2007). Learning-Induced Enduring Changes in Functional Connectivity among Prefrontal Cortical Neurons. *Journal of Neuroscience*, 27(4):909–918.

Balaguer et al., 2016. Balaguer, J., Spiers, H., Hassabis, D., and Summerfield, C. (2016). Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. *Neuron*, 90(4):893–903.

Botvinick et al., 2019. Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, 23(5):408–422.

Botvinick and Weinstein, 2014. Botvinick, M. and Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130480–20130480.

Botvinick, 2012. Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6):956–962.

Brendel et al., 2011. Brendel, W., Romo, R., and Machens, C. K. (2011). Demixed Principal Component Analysis. *Advances in Neural Information Processing Systems*, pages 2654–2662.

Buonomano and Maass, 2009. Buonomano, D. V. and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113–125.

Carnevale et al., 2015. Carnevale, F., de Lafuente, V., Romo, R., Barak, O., and Parga, N. (2015). Dynamic Control of Response Criterion in Premotor Cortex during Perceptual Detection under Temporal Uncertainty. *Neuron*, 86(4):1067–1077.

Chaisangmongkon et al., 2017. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., and Wang, X.-J. (2017). Computing by Robust Transience: How the Fronto- Parietal Network Performs Sequential, Category-Based Decisions. *Neuron*, 93(6):1504–1517.e4.

Collins and Koechlin, 2012. Collins, A. and Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLoS Biology*, 10(3):e1001293.

Daw et al., 2011. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215.

Domenech and Koechlin, 2015. Domenech, P. and Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, 1:101–106.

Doya, 1910. Doya, K. (1910). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10:732–739.

Frank, 2005. Frank, M. J. (2005). Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17(1):51–72.

Frank et al., 2004. Frank, M. J., Seeberger, L. C., and OReilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, 306(1):71–100.

Friston, 2005. Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

Gallego et al., 2017. Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, 94(5):978–984.

Hennequin et al., 2014. Hennequin, G., Vogels, T. P., and Gerstner, W. (2014). Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. *Neuron*, 82(6):1394–1406.

Histed et al., 2009. Histed, M. H., Pasupathy, A., and Miller, E. K. (2009). Learning Substrates in the Primate Prefrontal Cortex and Striatum: Sustained Activity Related to Successful Actions. *Neuron*, 63(2):244–253.

Houk, 1995. Houk, J. C. (1995). A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement. *Models of Information Processing in the Basal Ganglia MIT Press*, pages 249–274.

Johnson et al., 2018. Johnson, M., Townsend, J., Hawkins, P., and Maclaurin, D. (2018). JAX: Autograd and XLA. *Retrieved from https://github.com/google/jax*.

Kobak et al., 2016. Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, e10989:1–36.

Koechlin and Summerfield, 2007. Koechlin, E. and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235.

Kveraga et al., 2007. Kveraga, K., Ghuman, A. S., and Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition*, 65(2):145–168.

Li and Daw, 2011. Li, J. and Daw, N. D. (2011). Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction. *Journal of Neuroscience*, 31(14):5504–5511.

Lim and Goldman, 2013. Lim, S. and Goldman, M. S. (2013). Balanced cortical microcircuitry for maintaining information in working memory. *Nature Neuroscience*, 16(9):1306–1314.

Maclaurin et al., . Maclaurin, D., Duvenaud, D., Johnson, M., and Townsend, J. Autograd.

Mante et al., 2013. Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.

Nakahara et al., 2001. Nakahara, H., Doya, K., and Hikosaka, O. (2001). Parallel Cortico-Basal Ganglia Mechanisms for Acquisition and Execution of Visuomotor Sequences - A Computational Approach. *Journal of Cognitive Neuroscience*, 13(5):626–647.

Nicola and Clopath, 2017. Nicola, W. and Clopath, C. (2017). Supervised learning in spiking neural networks with FORCE training. *Nature Communications*, 8(1):1–15.

O'Doherty et al., 2004. O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304:452–454.

Pasupathy and Miller, 2005. Pasupathy, A. and Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433:873–876.

Qi and Constantinidis, 2012. Qi, X.-L. and Constantinidis, C. (2012). Variability of Prefrontal Neuronal Discharges before and after Training in a Working Memory Task. *PLoS ONE*, 7(7):e41053–8.

Rabinovich et al., 2008. Rabinovich, M., Huerta, R., and Laurent, G. (2008). Transient Dynamics for Neural Processing. *Science*, 321(5885):47–48.

Radulescu et al., 2019. Radulescu, A., Niv, Y., and Ballard, I. (2019). Holistic Reinforcement Learning: The Role of Structure and Attention. *Trends in Cognitive Sciences*, pages 1–15.

Rajan et al., 2016. Rajan, K., Harvey, C. D., and Tank, D. W. (2016). Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142.

Remington et al., 2018. Remington, E. D., Narain, D., Hosseini, E. A., and Jazayeri, M. (2018). Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron*, 98(5):1005–1019.e5.

Rich and Wallis, 2016. Rich, E. L. and Wallis, J. D. (2016). Decoding subjective decisions from orbitofrontal cortex. *Nature Neuroscience*, 19(7):973–980.

Sakai, 2008. Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, 31(1):219–245.

Salin and Bullier, 1995. Salin, P.-A. and Bullier, J. (1995). Corticocortical Connections in the Visual System: Structure and Function. *Physiological Review*, 75(1):1–48.

Samejima et al., 2005. Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, 310(5752):1333–1337.

Sarvestani et al., 2011. Sarvestani, I. K., Lindahl, M., Hellgren-Kotaleski, J., and Ekeberg, O. (2011). The arbitration–extension hypothesis: a hierarchical interpretation of the functional organization of the basal ganglia. *Frontiers in Systems Neuroscience*, 5(13):1–12.

Seo et al., 2012. Seo, M., Lee, E., and Averbeck, B. B. (2012). Action Selection and Action Value in Frontal-Striatal Circuits. *Neuron*, 74(5):947–960.

Shenoy et al., 2013. Shenoy, K. V., Sahani, M., and Churchland, M. M. (2013). Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience*, 36(1):337–359.

Song et al., 2017. Song, H. F., Yang, G. R., and Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6(e21492):1–24.

Stokes et al., 2013. Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, 78(2):364–375.

Summerfield et al., 2006. Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., and Hirsch, J. (2006). Predictive Codes for Forthcoming Perception in the Frontal Cortex. *Science*, 314:1311–1314.

Suri and Schultz, 1998. Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp Brain Res*, 121:350–354.

Sussillo and Barak, 2013. Sussillo, D. and Barak, O. (2013). Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks . *Neural Computation*, pages 1–24.

Sutskever, 2013. Sutskever, I. (2013). Training Recurrent Neural Networks. *PhD Thesis*, pages 1–101.

Sutton and Barto, 2018. Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning: An Introduction. *MIT Press, Cambridge, MA*.

Tsujimoto et al., 2008. Tsujimoto, S., Genovesio, A., and Wise, S. P. (2008). Transient Neuronal Correlations Underlying Goal Selection and Maintenance in Prefrontal Cortex. 18(12):2748–2761.

Verschure et al., 2014. Verschure, P. F. M. J., Pennartz, C. M. A., and Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130483–20130483.

Wallis et al., 2019. Wallis, J. D., Anderson, K. C., and Miller, E. K. (2019). Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411:953–956.

Wang et al., 2017. Wang, J., Narain, D., Hosseini, E. A., and Jazayeri, M. (2017). Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience*, 21(1):1–13.

Wang et al., 2018. Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):1–14.

Wood and Grafman, 2003. Wood, J. N. and Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, 4(2):139–147.

Yamins et al., 2014. Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–8624.

Yang et al., 2018. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. (2018). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):1–16.