How cortico-basal ganglia-thalamic subnetworks can shift decision policies to maximize reward rate

Jyotika Bahuguna[a,1], **Timothy Verstynen**[a,b,*], **and Jonathan E. Rubin**[b,c,*]

[a]Department of Psychology & Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America; [b]Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania, United States of America; [c]Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [*]Co-senior authors; [1]Current address: Université de Strasbourg, Laboratoire de Neurosciences Cognitives et Adaptatives (LNCA), CNRS, UMR 7364, Strasbourg, France

This manuscript was compiled on May 21, 2024

All mammals exhibit flexible decision policies that depend, at least in part, on the cortico-basal ganglia-thalamic (CBGT) pathways. Yet understanding how the complex connectivity, dynamics, and plasticity of CBGT circuits translates into experience-dependent shifts of decision policies represents a longstanding challenge in neuroscience. Here we used a computational approach to address this problem. Specifically, we simulated decisions driven by CBGT circuits under baseline, unrewarded conditions using a spiking neural network, and fit the resulting behavior to an evidence accumulation model. Using canonical correlation analysis, we then replicated the existence of three recently identified control ensembles (*responsiveness*, *pliancy* and *choice*) within CBGT circuits, with each ensemble mapping to a specific configuration of the evidence accumulation process. We subsequently simulated learning in a simple two-choice task with one optimal (i.e., rewarded) target. We find that value-based learning, via dopaminergic signals acting on cortico-striatal synapses, effectively manages the speed-accuracy tradeoff so as to increase reward rate over time. Within this process, learning-related changes in decision policy can be decomposed in terms of the contributions of each control ensemble, and these changes are driven by sequential reward prediction errors on individual trials. Our results provide a clear and simple mechanism for how dopaminergic plasticity shifts specific subnetworks within CBGT circuits so as to strategically modulate decision policies in order to maximize effective reward rate.

decision-making | value-based learning | cortico-striatal synaptic plasticity | drift diffusion model | control ensembles

A characteristic of nearly all mammals is the ability to flexibly shift how currently available evidence is used to drive actions based on past experiences (1). For example, feedback may be used to quickly shift between making exploratory decisions, where actions are sampled randomly or in order to gain information, and exploitative decisions, where actions are taken to maximize immediate rewards (2–4). Orthogonal to this exploration-exploitation dimension is a complementary choice about decision speed: actions can be made quickly or slowly depending on immediate goals and priorities (5). These shifts between fast or slow and exploratory or exploitative decision policies can be interpreted as different states of an underlying evidence accumulation process (6, 7), often captured by mathematical models such as the drift diffusion model (DDM; (8–12)). Any fixed values of parameters such as the drift rate (v; the rate of evidence accumulation during a single decision) and boundary height (a; the amount of evidence needed to trigger a decision) effectively represent a position on a manifold of possible decision policies that determine how both internal and external evidence combine to drive eventual actions (Figure 1, "WHAT" panel). The goal of learning is thus to converge to the position on this manifold of decision policies that optimally manages the speed-accuracy tradeoff for a given context (13).

This form of learning is managed, at least in part, by the cortico-basal ganglia-thalamic (CBGT) circuit, a distributed set of interconnected brain regions that can potentially influence nearly every aspect of decision-making (14–18) (Figure 1, "WHERE" panel). The CBGT circuit includes a collection of interacting basal ganglia pathways that receive cortical inputs and compete for control of an output region (predominantly the internal globus pallidus, GPi, in primates or the substantia nigra pars reticulata, SNr, in rodents) that impacts thalamocortical or superior collicular activity to influence actions (19–21). The balance of this competition is thought to map to a configuration of the evidence accumulation process (7, 22–26). Therefore, if behavioral flexibility reflects the *what* and CBGT circuits represent the *where* of flexible decision-making, then we are left with an open question of *how*: how do CBGT circuits achieve and control flexibility in decision policies during learning?

In prior work we showed how the computational logic of normative CBGT circuits can be expressed in terms of three low-dimensional subnetworks, called control ensembles, that each tune specific configurations of evidence accumulation parameters and reflect control over distinct dimensions of a decision policy (27). In theory, these control ensembles, dubbed *responsiveness*, *pliancy*, and *choice* (Figure 1, "HOW" panel), provide candidate mechanisms for controlling shifts in decision policies during learning. Here we illustrate how a single plasticity mechanism acting at the cortical inputs to the basal ganglia can, through network interactions, leverage the control ensembles to steer behavior during learning. To this end, we simulated a biologically-constrained spiking CBGT model that learns to select one of two actions via dopamine-dependent plasticity, driven by reward prediction errors, at the cortico-striatal synapses. We then implemented an upwards mapping approach (28), in which the behavioral features (decision times and choices) produced by the simulated CBGT network were modeled across stages of learning using the DDM (see (24, 27, 29)). Finally, we used various analytical approaches to replicate the existence of the low-dimensional control ensembles prior to learning and quantify how their influence levels change over the course of training. Our results

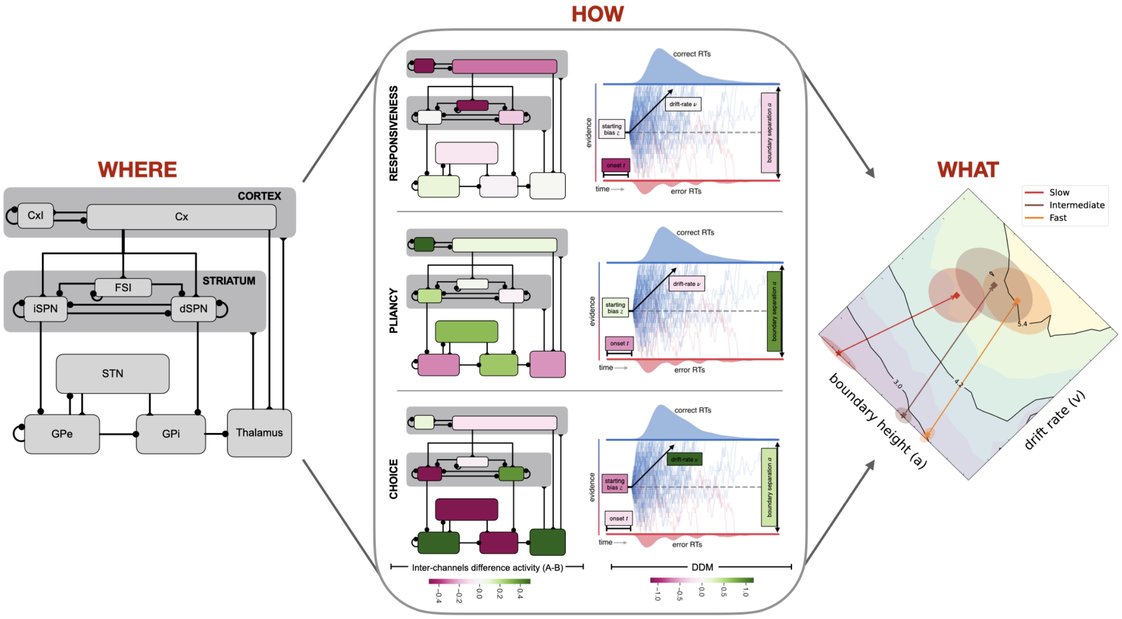

Fig. 1. Decision-making deconstructed. Most voluntary decision policies depend on the CBGT circuits (WHERE; left panel). This can be described at the algorithmic level by a set of parameters in a process model (e.g., the DDM) that drives an evidence accumulation process and determines the effective reward rate (WHAT; right panel contours), as well as other decision parameters. Control ensembles within CBGT circuits determine the relative configuration of decision policy parameters (HOW; middle panel) (27). What remains unclear is how learning drives changes in control ensembles that shift decision policies so as to maximize reward rate. Cx, cortical PT cells; CxI, inhibitory interneurons; FSI, fast spiking interneurons; d/iSPN, direct/indirect spiny projection neurons; STN, subthalamic nucleus; GPe, external globus pallidus; GPi, internal globus pallidus

show that value-based learning leads to a specific tuning of CBGT control ensembles in a way that maximizes the increase in reward rate across successive decisions.

Results

Feedback learning in CBGT networks maximizes reward rate.

Learning in the context of action selection involves finding an effective balance between the speed and accuracy of decisions (13). Here we consider a situation where an agent encounters a new environment for which it has no relevant prior experience or bias, so that the selection of all options is equally likely at first. In a simple two-choice bandit task, with one rewarded and one unrewarded option, this unbiased starting point would correspond to a 50% error rate. With learning it should be possible to make fewer errors over time, but exactly how this is achieved in practice depends on the decision policy that the agent adopts. For example, if the agent prioritizes speed over all else in its action selection, then its error rate will likely remain high. Conversely, by making sufficiently slow decisions, the agent may be able to achieve an extremely low error rate. The overall reward rate achieved by the agent depends on both decision speed and accuracy; intuitively this may be optimized for a fixed level of experience via some compromise between these two dimensions.

To understand how this optimization of speed and accuracy can emerge from CBGT circuits, we first simulated 300 instances of a spiking computational model of the CBGT pathways, each with a parameter set selected pseudorandomly from pre-determined parameter intervals that maintain the firing rates of the relevant cell types within known biological ranges (updated slightly from our past work (27); see Supporting Information Appendix, SI - Figure S1A). The networks performed a two-armed bandit task with deterministic reward feedback (i.e., the reward probability was 100% for the optimal choice and 0% for the suboptimal one). Learning was implemented with dopamine-dependent plasticity at the cortico-striatal synapses, where the magnitude of the phasic dopamine response was based on reward prediction errors (for details see (30)). We fit the reaction times (RT) and choice probabilities of each network with a hierarchical version of the DDM (31, 32). The DDM provides an intuitive framework for mapping behavioral responses to an evidence-accumulation decision policy that can be described by only a few parameters (8). After each predetermined step in learning (2, 4, 6, and 15 trials with plasticity on), we would freeze the network by turning off plasticity, simulate 300 trials to generate an RT distribution and choice probabilities, and fit the DDM to these behavioral measures. After these probes, learning was turned back on and the task progressed. This process yielded an effective trajectory in the DDM parameter space.

Figure 2 shows the average trajectories of three groups of networks on a manifold defined by two parameters of the DDM, drift rate (v) and boundary height (a). For each v and a we also estimated the average RT (Figure 2A), accuracy (Figure 2B) and reward rate (Figure 2C). The three groups represent a tertiary split of the full set of simulated networks into fast (short RT, orange), intermediate (medium RT, brown), and slow (long RT, red) groups, based on their initial RT values (Figure S1B). We implemented this split to determine whether decision policy adjustments due to learning were influenced by initial biases in the networks. Despite their initial speed differences, all three network classes showed chance level performance before plasticity (Figure S1C) and converged to similar regions of the (v, a) space with learning (Figure 2, shaded ellipses). A comparison of behavioral measures and DDM parameters before and after plasticity is presented in Figure S2.

These trajectories clearly demonstrate that our CBGT network can learn from simple dopaminergic feedback at cortico-striatal synapses. But what exactly is the objective being maximized by the network? To test this, we compared the change at each step of learning to the predicted direction that the network would take if it were maximizing one of three possible behavioral objectives: speed, accuracy, or reward rate. These predicted directions are illustrated as blue vectors in Figure 2A-C, reflecting steps from each initial point that are in the direction of the gradient of each objective (i.e., the direction of maximal change, which lies orthogonal to the contours, shown with the same length as the vector representing the actual network evolution at the first step of learning in each case). Analysis of the trajectories in Figure 2A reveals that while plasticity decreases RTs with learning, the angles of the learning trajectories do not align with the optimal directions for maximally reducing RT. Similarly, the network trajectories do not align with the vectors that would be expected if they were maximizing accuracy alone (Figure 2B). In contrast, the average trajectories along the reward rate manifold (Figure 2C) were closest to the optimal direction. Moreover, the rate of increase in reward rate was similar regardless of the network's initial speed bias.

To quantify the alignment of observed network trajectories to the expected directions of maximal change, we calculated the cosine distance between the observed vector and the optimal vector, normalized to the observed vector's length, at each learning step. While there is substantial variability across networks (Figure 2D), there was a consistent effect of objective type on network fits (F[3813, 2]=47.2, p<0.0001). Fits to the reward rate trajectories (cosine distances averaged over all plasticity stages for each network) were consistently better than to either RT (t(299)=13.22, p<0.0001) or accuracy (one-sample t(299)=8.75, p<0.0001) trajectories. This effect held regardless of a network's initial bias (Figure S3). Thus, our biologically detailed model of the CBGT circuit can effectively learn to maximize reward rate by managing the speed-accuracy tradeoff during the evidence accumulation process via dopaminergic plasticity at the cortico-striatal synapses.

Low-dimensional control ensembles that map to general decision policies. The CBGT network and DDM are, respectively, implementation-level and algorithmic-level descriptions of the evidence accumulation process that guides goal-directed behavior. We have previously shown that there is a low-dimensional, multivariate mapping between these two levels of analysis in the absence of learning (27). Here we set out to replicate this observation with the CBGT parameter sets used in the current study, with the aim of analyzing their contributions to the dopaminergic learning process. For this step, we considered two aspects of activity within each CBGT population: global activation across the two action representations (sum of the activity in that region, across both channels; Σ) and bias towards one action representation (difference in activity within each region, across the action channels; Δ). Using canonical correlation analysis (CCA), we captured the low-dimensional components that maximally correlate variation in CBGT activity with variation in DDM parameters. This analysis identified three such components (Figure 3). We refer to these low-dimensional components as *control ensembles.*

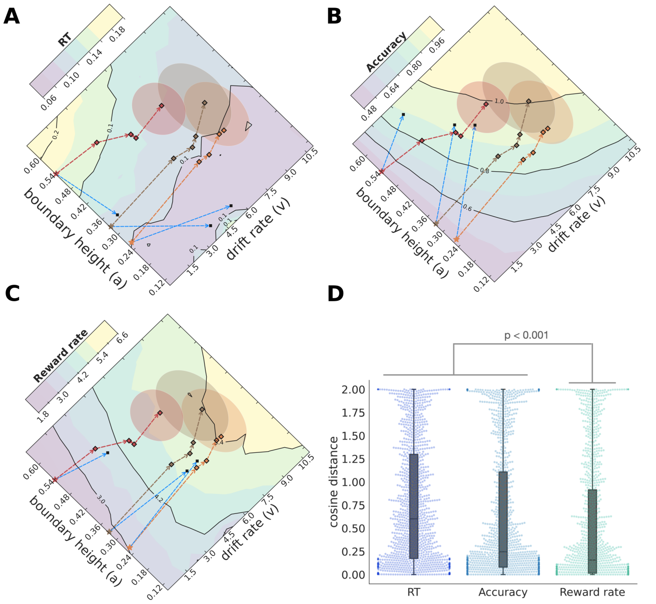

Fig. 2. Dopamine-dependent cortico-striatal plasticity drives CBGT networks in the direction of reward rate maximization. **(A)** The evolution of RTs achieved by a DDM fit to CBGT network behavior, projected to (v, a)-space. The orange (fast), brown (intermediate) and red (slow) stars represent the average starting positions of the three groups of networks with different initial decision speeds. The squares indicate the evolution of each network group over the plasticity stages, which converge after 15 trials (shaded ellipses). The yellow (purple) colors represent high (low) RTs. The network trajectories do not evolve in the direction that would be expected to minimize the RTs (e.g., optimal direction shown in blue from the initial position of all three speed groups). **(B)** The yellow (purple) colors represent high (low) accuracy. The networks evolve towards increasing expected accuracy but not in an optimal fashion (trajectories vs. blue arrows). **(C)** The yellow (purple) colors represent high (low) reward rate. The network evolution aligns closely with the direction that maximizes the reward rate (blue arrows). **(D)** The cosine distances calculated for every network at each plasticity stage for RT, accuracy and reward rate are shown as distributions.

The three control ensembles identified by our analysis nearly perfectly replicate our prior work (27), where they are described in more detail (see also Section *Upward mapping*). Thus we kept the labels *responsiveness*, *pliancy*, and *choice* ensembles for the first, second, and third components recovered, respectively. The recovered components are shown in both CBGT and DDM parameter spaces in Figure 3 (right panels). The responsiveness component describes the agent's sensitivity to evidence, both in terms of the delay before the agent starts to accumulate evidence (t) and how significantly the presence of evidence contributes to achieving the decision threshold (a). The dominant features of CBGT activity that vary along the responsiveness control ensemble loadings are a global inhibitory signal, including fast-spiking interneuron (FSI) and overall internal globus pallidus (GPi(Σ)) activity, as well as overall excitatory and inhibitory cortical activity (Cx(Σ), CxI). Because the CBGT and DDM loadings that emerge from the CCA have the same sign (all negative), they imply that a *decrease* in the weighted activity of the loaded cells corresponds to an *decrease* in t and a and hence to a *increase* in overall responsiveness.

The pliancy component refers to the level of evidence that must be accumulated before committing to a decision. As with responsiveness, pliancy loads mostly on a and t, but now with

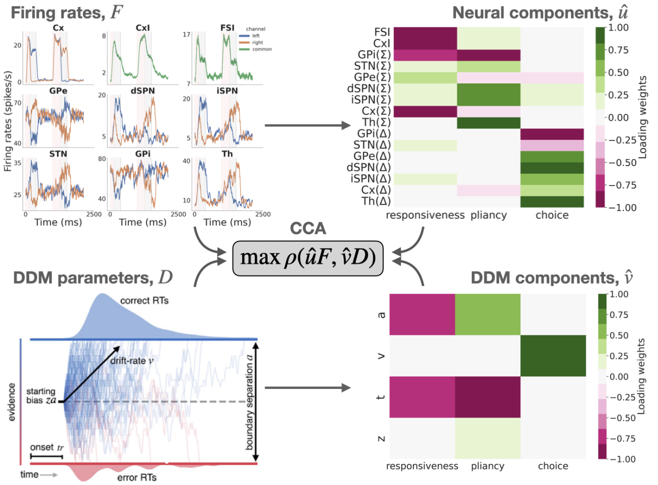

Fig. 3. Canonical correlation analysis (CCA) identifies control ensembles (cf. (27)). Given matrices of average firing rates, F (both summed rates across channels, Σ, and between-channel differences, Δ), and fit DDM parameters, D, derived from a set of networks at baseline (left panels), CCA finds the low dimensional projections, $\hat{u}$ for firing rates and $\hat{v}$ for DDM parameters (right panels), which maximize the correlation, ρ, between the projections $\hat{u}F$ and $\hat{v}D$ of F and D.

opposing signs for these two loadings, corresponding to the idea that even though an agent is attentive to evidence (small t), it requires significant evidence to reach its threshold (large a). The CBGT activity features that characterize pliancy are the overall engagement of the BG input nodes (i.e., global dSPN and iSPN activity, with a smaller STN contribution), as well as total GPi and thalamic activity, with opposite loadings to each other. For the pliancy component, a change in the activity consistent with the cell type loadings (e.g., increase in SPN activity) corresponds to a decrease in overall pliancy (e.g., increase in a).

Lastly, the choice component represents the intensity of the choice preference and is reflected largely in the drift rate (v) and the neural correlates of competing choice representations in the CBGT (i.e., difference in activity across the two action channels within each BG region). A change in activity consistent with the cell type loadings (e.g., greater difference in dSPN activity between the two channels) corresponds to a stronger commitment towards the more rewarded option (i.e., larger v).

In summary, each CBGT control ensemble can be interpreted as specifying a coordinated collection of changes in CBGT neural activity levels that can, in theory, most effectively tune a set of decision policy parameters (captured by the DDM). Now that we have delineated the control ensembles embedded within the CBGT network (cf. (27)), we are ready to consider how dopamine-dependent plasticity regulates their influence in a way that collectively drives decision policies to maximally increase reward rate.

Cortico-striatal plasticity drives control ensembles during learning. Our analysis of the CBGT network behavior (Figure 2) shows that dopamine signaling at the cortico-striatal synapses is enough to elicit changes in the evidence accumulation process that maximize reward rate. This observation suggests that there are emergent driver mechanisms, originating from cortico-striatal synaptic changes, that tune the control ensembles in a way that achieves this outcome. That is, if each control ensemble represents a knob to tune an aspect of the decision policy, then a driver mechanism selects a set of adjustments of the knobs that yields an overall decision policy selection. We next set out to identify these emergent drivers.

As a first step, to quantify the modulation of CBGT activity after plasticity, we calculated the principal components of the change in firing rates of all 300 networks, before and after plasticity. The first 5 of these components collectively explain more than 90% of the observed variance (Fig. S4A, thick blue line marked "All"). The loading weights (Fig. 4A) show that the first and third components reflect the global activity of the CBGT nuclei. The second, fourth and fifth components relate more strongly to the bias towards one option, with predominant loadings on differences in rates across channels in certain CBGT regions. Together, these components represent the collection of changes in firing rates that result from learning-related changes at the cortico-striatal synapses.

We next calculated the matrix S of weighting factors (*drivers*) for the firing rate components, describing what combination of adjustments to the control ensembles best accounts for the associated firing rate changes (Fig. 4B; for full description of this approach see Methods subsection *Modulation of control ensembles by plasticity*). To interpret the drivers of control ensemble influence (Fig. 4B), it is important to note that positive (negative) coefficients correspond to changes in control ensemble activity in the same (opposite) direction as indicated by the loadings in Fig 3. The first driver corresponds to a large amplification of the responsiveness control ensemble, and hence a decrease in various forms of global inhibition in the CBGT network (overall GPi, FSI and CxI activity), along with a boost to the choice control ensemble, and hence increased bias towards the rewarded choice (differences in activity across CBGT channels). The second driver has a strong negative weight on the choice and a positive weight on the pliancy control ensemble. The third, fourth and fifth drivers feature weaker effects, with small modulations of all three control ensembles. Based on this analysis across all of the networks, the overall modulation of the control ensembles due to plasticity, calculated as the weighted sum over all drivers (weighted by the % of variance explained by each PC), is shown in Supp Figure S4B. All three control ensembles end up being boosted, meaning that, to varying extents, the activity measures that comprise these ensembles change in the directions indicated by their loadings in Fig. 3. In this way the general trend is for the CBGT networks to become more responsive, yet less pliant, which together amount to an earlier onset of evidence accumulation without much change in boundary height, and exhibit more of an emergent choice bias.

Because of the difference in decision policies across the fast, intermediate, and slow networks, we recomputed the drivers separately for each network type. This was done by considering the firing rate differences (ΔF) and calculating the S loadings for fast, intermediate, and slow networks separately (see Methods - section *Modulation of control ensembles by plasticity*). The explained variance for the three network types are shown in Supp Figure S4A, and their corresponding PCs and goodness of fits are shown in Supp Figure S5. As expected, the drivers showed variability across the network types (Fig. 4C). The driving factor corresponding to responsiveness is negative

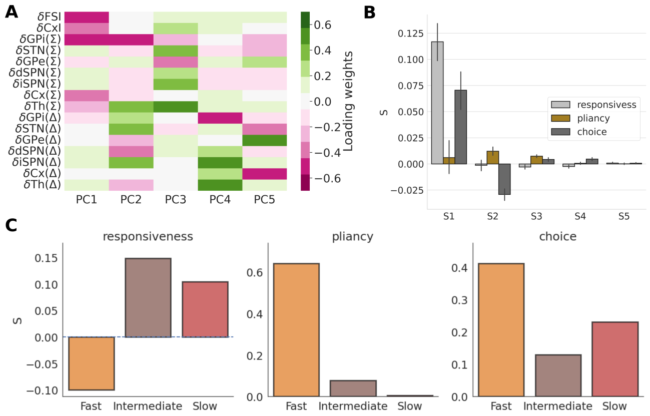

Fig. 4. Corticostriatal synaptic plasticity results in increased pliancy and choice ensemble activity in all CBGT networks; however, the sign of the responsiveness change depended on network class. **A)** The loading weights of the first 5 PCs of firing rate changes from before to after plasticity pooled for all networks. **B)** The drivers (columns of S), which quantify the modulation of control ensembles (responsiveness, pliancy, choice) that capture each PC (pooled for all network classes). **C)** The variance-weighted combinations of drivers for each control ensemble, combined separately for the three network classes (fast, intermediate and slow).

for fast networks, while remaining positive for the others. The pliancy and choice factors were positive for all three networks, but pliancy was by far the largest for fast networks and quite small for the other two network types. Referring to the DDM parameter changes associated with changes in control ensemble loadings (Fig. 3), we see that the decrease in responsiveness and strong increase in pliancy for fast networks would both promote an increase in boundary height, a. This aligns with the fact that, of the three network types, only fast networks show an increase in a over the course of learning (Fig. 2, Supp. Fig. S6). Overall, we see that the specific way that plasticity adjusts the weighting of the control ensembles to drive changes in decision policies depends on the current state of the network. Since plasticity results from the sequence of decisions and rewards that occur during learning, we next investigate more directly how decision outcomes lead to this dependency.

The influence of feedback sequences on control ensembles. In the previous section, we described the overall effects of corticostriatal plasticity on control ensemble tuning. To build from there, we next analyzed the early temporal evolution of these effects by focusing on the initial two learning trials. Specifically, we examined the modulation of the control ensembles for different combinations of successes (i.e., rewarded trials; R) and failures (i.e., unrewarded trials; U) achieved by the first two consecutive choices. For this analysis, we implemented our usual DDM fitting process followed by CCA for networks that were frozen (i.e., with plasticity switched off) after two trials, and we grouped the results based on the sequence of choice outcomes. The drivers (combined columns of S) for each sequence of outcomes, U-U, U-R, R-U and R-R, are shown in Fig. 5A.

First, consider the case of networks that receive no rewards (U-U). Here we infer that the boundary height, a, increases, due to a simultaneous decrease in driving of the responsiveness ensemble and increase in driving of the pliancy ensemble, both of which result in a boost of the boundary height. In addition, driving of the choice ensemble is reduced. Thus, two consecutive unsuccessful trials yields an overall increase in the degree of evidence needed to make a subsequent decision by simultaneously increasing the boundary height and decreasing the drift rate. Moreover, slow networks encounter U-U outcomes more often than other network classes in the first two trials (Supp. Table 1), which presumably constrains the increase in responsiveness and choice seen in these networks during learning (Fig. 4C). On average, however, fast networks make more mistakes than the other networks. This result, which we can display graphically in terms of the proportion of unrewarded trials, or mistakes, encountered after the first two plasticity trials (Fig. S6D), likely explains the negative loading for responsiveness and high positive loading for pliancy for fast networks shown in Fig. 4C.

In contrast, two consecutive successful trials (R-R, far right of Fig. 5A) produce largely the opposite effect. The influences of the responsiveness and choice ensembles increase, resulting in lower onset time and boundary height along with an increase in the drift rate. This coincides with a weak change in pliancy. As a result, in the R-R case, the decision policy is tuned to include a decreased degree of evidence needed to make subsequent decisions.

Not surprisingly, the two mixed combinations of outcomes (U-R, R-U) have largely similar effects on the responsiveness and pliancy ensembles, regardless of the order of outcomes. In both cases responsiveness increases and pliancy decreases, resulting in less overall evidence needed to trigger a decision (by shrinking the boundary height, without much change in the onset time). However, when the first trial is unsuccessful (U-R) the influence of the choice ensemble decreases, while it increases when the first trial is successful (R-U). Indeed, looking at the progressive change in the choice ensemble across the four unique sequences of trials, it appears that early success (i.e., reward in the first trial) boosts the choice ensemble influence while early failure (i.e., unrewarded first trial) does the opposite. When these combined drivers are recomputed separately for each network class, the learning-induced modulations of the ensembles follow the same general trend (Supp. Figure S7), with quantitative details depending on the network class.

The preceding analysis shows how the relative contributions of the control ensembles to the evidence accumulation process depend on trial outcomes. What are the results of these changes on the performance of the network? To illustrate these effects, we plot the distribution of changes in reward rates associated with each set of outcomes and separate by network types in Fig. 5B. Although all distributions are generally positive, there is significant variation in reward rate changes across the different feedback sequences (F(619 , 3) = 274.2, p<0.0001). The reward rate also varies significantly with the network type (F(619, 2) = 50.3, p<0.0001), and the interaction term between network types and feedback sequences is significant as well (F(619, 6)=3.5, p = 0.002). Compared to all other conditions, the networks that made two consecutive unsuccessful choices (U-U) yielded the smallest changes in reward rates (values of all network types pooled together, all two-sample t(336) > -19.11, all p<0.0001). The two mixed feedback conditions (U-R, R-U) had substantially higher growth in reward rates than the condition with two rewarded trials (R-R; all t(404) > 8.38, all p<0.001), perhaps

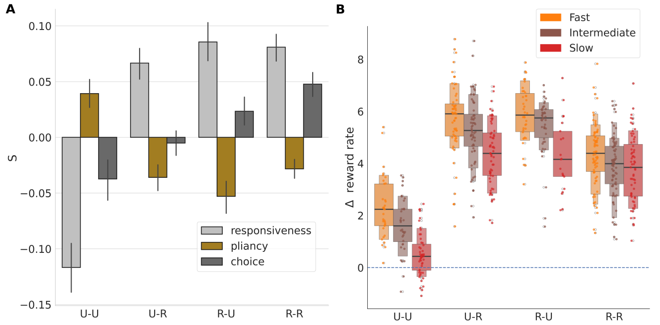

Fig. 5. Suboptimal and optimal choices modulate control ensembles in opposite directions. **A)** The modulation of control ensembles associated with various reward sequences encountered in two initial trials with corticostriatal plasticity. U represents "Unrewarded" and R represents "Rewarded" trials. **B)** The reward rate changes obtained by simulation of networks with synaptic weights frozen after various reward sequences occurred on two initial trials.

because R-R sequences were more likely in networks that already had high reward rates. In all cases, the trend was for faster networks to achieve greater increases in reward rate. As expected, the impact of feedback sequences on reward rate is associated with underlying changes in both accuracy (Fig. S8A) and decision speed (Fig. S8B).

Discussion

Adaptive behavior depends on flexible decision policies (**what**), driven by CBGT networks (**where**) that shift their activity in order to maximize reward rate by coordinated adjustments of a set of underlying control ensembles (**how**; Fig. 1). In this work, we focused on the **how** part of this process, using an upward (in abstraction) mapping between a biologically realistic model of CBGT pathways and the DDM to illustrate the complex, low-dimensional structure of the CBGT subnetworks that modify decision policies (Fig. 3). Specifically, we recapitulated recent results (27) showing that the three main CBGT control ensembles of decision-making represent *responsiveness*, *pliancy*, and *choice* (Fig 3) and serve to regulate the evidence accumulation process. We then showed how driver mechanisms tune these control ensembles strategically during learning (Fig. 4 & 5) in order to maximize reward rate. Moreover, although they all optimize the same quantity (reward rate), we find that networks modulate the control ensembles differently depending on their *a priori* decision policy (fast, intermediate, or slow). While all networks increase responsiveness and choice to varying extents, fast networks alone decrease responsiveness (Fig. 4C) and correspondingly increase boundary height (a; Fig. S5A). Put together, our results highlight the dynamic and coordinated way that subnetworks within CBGT circuits can regulate adaptive decision-making through simple dopaminergic plasticity at the cortico-striatal synapses.

Perhaps the most surprising aspect of this theoretical analysis is the sophisticated adjustments that emerged from a simple plasticity mechanism on just one class of CBGT synapses. Dopaminergic learning at the cortico-striatal synapses was sufficient to push our naive networks from an exploratory decision policy to an exploitative policy that effectively managed the speed-accuracy trade off by maximizing average reward rate (Fig. 2). This behavior was also recently observed in rats performing a perceptual learning task (33), suggesting that such reward rate maximization is a natural behavior in many, if not all, mammals. The rewards in our task that drove learning were based only on the accuracy of each selection. So, how is it possible that rewards based only on accuracy lead to an optimization of reward rate? The answer to this question lies in the architecture of the CBGT circuits. Although the synaptic plasticity in our model occurs only at the cortico-striatal synapses, the changes in activity that result from this plasticity ripple throughout the entire CBGT network, based on the synaptic coupling among populations that the network includes. An emergent result from our simulations is that these cascading effects produce the subsequent reduction of decision times, even without any reward incentive that explicitly depends on speed. As a result, the model tends to act more slowly in the early phases of learning, but increases accuracy and speeds decisions as learning progresses. This is similar to behavioral observations in rodents (33, 34), non-human primates (35), and humans (36, 37). Our results suggest that this complex behavior is a natural consequence of dopamine-dependent plasticity at the cortico-striatal synapses together with the architecture of the CBGT circuit.

Here we decomposed the circuit-level effects of plasticity that underlie adaptive reward rate maximization in terms of varying levels of learning-related drives on a set of control ensembles. Based on the relation of the control ensemble loading to evidence accumulation parameters (Fig. 3), the effective learning-related changes result in shorter decision onset delays, higher rates of evidence accumulation, and variable changes in decision threshold as learning progresses (Fig. S6). On the shorter timescale of consecutive trials, each possible set of reward outcomes induces a specific adjustment of control ensembles in a way that increases subsequent accuracy and reward rate (Fig. 5, Fig. S8). Interestingly, but perhaps not surprisingly, having mixed feedback (one rewarded and one unrewarded trial) resulted in more effective reward rate maximization than two consecutive rewarded trials, consistent with past results (and intuition) on the benefits of exploration for effective learning (38, 39). It is, however, important to note that cortico-striatal plasticity may explain only a part of the decrease in decision speed seen in experiments, with additional reductions that result from an agent's increased confidence in the outcomes of its decisions (increased certainty) deriving from other information sources (40). Moreover, an experimental paradigm that requires learning an explicit minimization of decision times may reveal other novel CBGT control ensembles, apart from those that we report here.

A reasonable question at this point is whether the control ensembles that play a crucial role in learning in our simulations exist in real CBGT circuits. Directly recovering these ensembles would necessitate simultaneous *in vivo* recording of nine distinct cell populations during a learning task. This is currently outside the scope of available empirical technology. Nonetheless, a review of the current literature reveals piecemeal indications of the existence of these control ensembles. For example, the predominant loadings in the responsiveness ensemble in our CBGT model corresponds to decreases in FSI, cortical, and overall GPi activity. The increase in responsiveness associated with learning in intermediate and slow networks in our model therefore matches the suppression of activity in the subpopulation of striatal FSIs that was observed after learning in non-human primates (41). Interestingly, experiments have also found evidence for an earlier onset of activity in striatum with the progression of learning in non-human primates (42), consistent with the decrease in onset time t that arises via the learning-induced increase in drive of the responsiveness or pliancy ensemble in all network classes in our model.

The pliancy ensemble is associated with the onset time and boundary height parameters, but with opposing loadings. Thus, an increase in activity of the pliancy ensemble corresponds to an earlier onset of evidence accumulation but with a higher boundary height. This places an emphasis not on the collection of evidence itself, but on the agent's willingness to be convinced by this evidence. It has been shown that an increase in the conflict between action values is associated with an increase in global STN activity (43–45), which is consistent with a strengthened driving of our pliancy ensemble that results in a higher decision threshold. Also, because our simulations show an increase in efficacy of the pliancy ensemble with value-based learning (Fig 4C) for fast and intermediate networks, we predict that the overall level of striatal SPN activity will increase as learning progresses, while that in GPi will decrease. The predominant contributions of this effect are predicted to occur in response to unrewarded trials (Fig 5A). Consistent with this idea, past studies have shown such increases in striatal activity in association with learning (46). Related findings have been interpreted as being potentially linked to increased attentiveness to a task (47) or increased motivation (48, 49). Both effects are consistent with the lowering of onset time associated with our pliancy ensemble. Interestingly, increases in striatal activity, as measured via fMRI, have been found to be beneficial for learning in adolescents (50), which our results suggest could relate to enhanced learning from mistakes.

Finally, the choice ensemble is strongly associated with drift rate. The CBGT components contributing to this ensemble include the differences across action representations in both dSPN and iSPN populations. Consistent with this relationship, single unit activity in dorsal striatum has been shown to reflect the rate of evidence accumulation and consequently preference for a specific response to a stimulus (51). At the macroscopic level, we recently found that the competition between action representations in CBGT circuits, measured with fMRI, is indeed reflected in the drift rate in humans (7). At the causal level, a recent study with patients suffering from dystonia showed that deep brain stimulation (DBS) in the GPi increased the likelihood of exploratory behavior, which was encoded as decrease in the drift rate in an DDM-type model (17). Whether DBS increases or decreases the output of its target area is a matter of controversy (52–54); however, based on the loadings in the choice ensemble, we would predict that the decrease in drift rate aligns with activity becoming more similar across GPi neurons in different channels, which would be a natural result if DBS affected all channels similarly.

Taken all together, the results in this paper show how the low-dimensional substructure of CBGT circuits can implement environmentally appropriate changes in behavior during learning by tuning specific aspects of the evidence accumulation process that, in turn, determine the current state of a decision policy. Importantly, dopamine-dependent synaptic plasticity at the cortico-striatal synapses, mediated by choice-related reward signals, adjusts the activity of these control ensembles in a strategic and coordinated way that improves accuracy while reducing decision times so as to maximize the increase of reward rate. These results not only align with previous empirical observations, but also make explicit predictions that can be the focus of future experimental work.

Materials and Methods

CBGT network. The CBGT network model is a biologically constrained spiking neural network including neural populations from the striatum (dSPNs, iSPNs and FSIs), globus pallidus external segment (GPe), subthalamic nucleus (STN), globus pallidus internal segment (GPi), thalamus and cortex (excitatory and inhibitory components). For a two-choice task, each choice representation is implemented as a “channel” (21, 24, 27, 55), so the model includes two populations of each type except FSIs and inhibitory cortical neurons, which are shared. The cortico-striatal projections to both dSPNs and iSPNs are plastic and are modulated by a dopamine-dependent spike timing dependent plasticity rule (29, 56, 57). On a trial, a choice is selected if the firing rate in the thalamic population within its action channel reaches 30 Hz before the rate of the other thalamic population hits that level. The complete details of this network can be found in our methods paper (30).

Characterization of networks before plasticity. In our previous work, we identified control ensembles based on extensive simulation of the CBGT network with each of 300 parameter sets selected using Latin hypercube sampling from among the ranges of synaptic weights that maintained biologically realistic firing rates across all populations (27). In that work, in which no learning occurred, however, the cortico-striatal projections to the choice representations (channels) were considered to be independent. Hence, some sampled network configurations were biased towards one of the choices. Because we study the evolution of the control ensembles under plasticity in this work, we started with completely unbiased networks. Hence we resampled the networks from the joint synaptic weight distribution using genetic algorithms (see below) and isolated 300 networks that produced firing rates of all CBGT populations within the experimentally observed ranges. The firing rate distributions are shown in Supp Fig S1A. The networks before plasticity showed a diversity of reaction times (RTs, Supp Fig S1B). The RT distribution was divided into 3 equal tertiles and used to define "fast" (orange), "intermediate" (brown) and "slow" (red) networks. All of the networks before plasticity showed chance levels of accuracy (Supp Fig S1C).

Genetic algorithms. The DEAP library (58) was used to run a genetic algorithm (GA) designed to sample networks with parameters from the ranges used previously (27). Two additional criteria were used for the optimization function of the GA, namely (a) the network should produce trial timeouts (when no action was selected within 1000 ms) on fewer than 1% of trials, and (b) the network should be cortico-basal-ganglia driven; that is, the correlation between cortical activity and striatal activity should be positive. The first criterion ensured that we had ample choices included in the data, as needed to appropriately fit the DDM parameters (timeouts are dropped before fitting the DDM parameters). The second criterion ensured that the networks did not operate in a cortico-thalamic driven regime, in which cortical inputs alone directly pushed thalamic firing over the decision threshold.

The range for each parameter specified in past work (27) was divided into 30 bins and this grid was sampled to create populations. The indices of each bin served as a pointer to the actual values of the parameters in the ranges considered. The GA uses these indices to create, mate and mutate the populations. This ensures that the values of parameters remain within their specified ranges. For example, suppose that parameter A has range (-2.0,2.0) and parameter B has range (-0.3,1.0) and these ranges are each divided into 5 bins. The grids for parameters A and B will be:

$$A_{grid} = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \end{pmatrix}$$
$$B_{grid} = \begin{pmatrix} -0.3 & 0.025 & 0.35 & 0.675 & 1 \end{pmatrix}.$$

If individual population members have indices $ind_1 = (0\ 1)$ and $ind_2 = (4\ 0)$ for (A, B), then they have $(A, B) = (-2, 0.025)$ and $(A, B) = (2, -0.3)$, respectively. Supposed that the individuals mate by crossing over the 1st and 2nd elements. Then $ind_3 = (4\ 1)$ with parameter values $(2, 0.025)$ and $ind_4 = (0\ 0)$ with parameter values $(-2, -0.3)$. The individuals ind_3 and ind_4 are included in the next iteration of evolution.

New individuals created from mating were used to overwrite the original individuals that were mated together(*cxSimulatedBinary*). The individuals could also mutate by shuffling the indices of the attributes (*mutShuffleIndexes*) with a probability of 0.2. After a round of mating and mutation, tuples of two values for each individual, namely the % of timeouts and the Pearson's correlation coefficient between cortical and striatal activity, were compared to select the individuals for the next round of evolution. The selection algorithm that was used was tournament selection (*selTournament*) of size 3, which picked the best individual among 3 randomly chosen individuals, 10 times, where 10 is the size of the population of networks in every iteration of the GA. During every iteration, any network configuration that met the criteria (a) and (b) above was saved as a correct solution. The GA was run for 2000 iterations or until 300 solutions were found, whichever was sooner. Post hoc, we confirmed that the firing rates of the members of the final, selected population remained within the originally targeted ranges (Figure S1).

Upward mapping. The DDM parameters and activity of the CBGT nuclei for our 300 network configurations, before plasticity, were used to identify CBGT control ensembles through canonical correlation analysis (CCA), as was also done in our previous work (27) and is illustrated in Fig 3. The CCA scores were calculated using k-fold validation (k=4), where the 300 networks were divided into groups of 4 (75 networks each) and a CCA score was calculated for each of the groups. The CCA scores for actual data were compared with a shuffled version of data (firing rates and DDM components for 300 networks) and the set of components giving rise to the maximum CCA score, which we found to include three elements as in our previous work (27), were selected.

Modulation of control ensembles by plasticity. We used a single approach to compute a set of effective drivers of the control ensembles either from the full collection of CBGT networks or from one of the network subtypes (fast, intermediate, or slow) that we considered. Let $X \in \{\texttt{all}, \texttt{fast}, \texttt{intermediate}, \texttt{slow}\}$ denote the class of networks being used. From the set of vectors of changes in CBGT firing rates computed by subtracting firing rates before plasticity from those after plasticity (ΔF_X), we extracted 5 principal components (PCs) that together explain at least about 90% of the variance (Fig. 4A, Supp Figure S4A). ΔF_X was then projected onto these 5 PCs to form the target matrix P_X. Specifically, we computed

$$P_X = (\Delta F_X)V_X \qquad [1]$$

where the 5 PCs comprise the columns of V_X. Note that P_X is an n by 5 matrix, where n is the number of firing rate data vectors used. ΔF_X was also projected onto the three control ensemble components obtained from the full collection of baseline networks before plasticity, via the mapping

$$C_X = (\Delta F_X)U \qquad [2]$$

where the components of the 3 control ensembles form the columns of U, such that C_X is an n by 3 matrix. Finally, we found the least squares solution S_X, representing the element in the range of C_X that is closest to P_X, from the normal equation

$$S_X = (C_X^T C_X)^{-1} C_X^T P_X. \qquad [3]$$

The least squares solution S_X is a 3×5 matrix independent of n. The columns of $S_{\texttt{all}}$ are displayed in Fig. 4B. The sums of the columns of the appropriate S_X, each weighted by the percent of variance explained, comprise Figs. 4C and S7 ($X = \texttt{fast}$, $X = \texttt{intermediate}$, and $X = \texttt{slow}$), as well as Figs. 5A and S4B ($X = \texttt{all}$).

Reward rates. The reward rate was calculated as:

$$\begin{aligned} RR &= \frac{1-\text{p(err)}}{DT+T_0} \\ &= \frac{\text{accuracy}}{RT} \end{aligned}$$

where p(err) denotes the error rate and where the reaction time, RT, is the sum of the decision time, DT, and the additional non-decision time that arises within each trial, T_0, which in our analysis is ascribed to the onset delay represented by the DDM parameter t.

Plasticity stages. The effect of plasticity on the network was studied at four stages: a) after 2 trials of plasticity, b) after 2 additional trials (total 4) of plasticity, c) after 2 more additional trials (total 6) of plasticity, d) after 9 additional trials (total 15) of plasticity. The state of the network was frozen at each of these stages by suspending the plasticity, so that we could use the frozen network to perform probe trials. The choices and reaction times from the probe trials were used to calculate DDM parameters and reward rate distributions for each stage of plasticity, based on upward mapping and CCA, and thus to generate the trajectories in Fig. 2, the time courses in Fig. S6, and the 2-trial results in Figs. 5, S7, and S8.

Data sharing. The network codebase utilized in this study can be found on our GitHub repository and accessed at https://github.com/CoAxLab/CBGTPy/blob/main. Detailed installation instructions and a comprehensive list of implemented functions can be found in the README.txt file within the repository. All datasets generated and analyzed during the course of this research, along with a demonstration demo will be openly available on GitHub at https://github.com/jyotikab/CBGT_maximize_RR.

ACKNOWLEDGMENTS. We thank all members of the exploratory intelligence group for their helpful comments on the manuscript. JB is supported by ANR-CPJ-2024DRI00039. TV, JB and JER are partly supported by NIH awards R01DA053014 and R01DA059993 as part of the CRCNS program. JER is partly supported by NIH award R01NS125814, also part of the CRCNS program.

1. JI Gold, MN Shadlen, The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
2. JD Cohen, SM McClure, AJ Yu, Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**, 933–942 (2007).
3. K Mehlhorn, et al., Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* **2**, 191 (2015).
4. RC Wilson, E Bonawitz, VD Costa, RB Ebitz, Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences* **38**, 49–56 (2021).
5. JT Dudman, JW Krakauer, The basal ganglia: from motor commands to the control of vigor. *Current opinion in neurobiology* **37**, 158–166 (2016).
6. K Bond, K Dunovan, A Porter, JE Rubin, T Verstynen, Dynamic decision policy reconfiguration under outcome uncertainty. *eLife* **10**, e65540 (2021).
7. KAM Bond, et al., Competing neural representations of choice shape evidence accumulation in humans. *eLife* **12**, e85223 (2023).
8. R Ratcliff, A theory of memory retrieval. *Psychological Review* **85**, 59–108 (1978).
9. R Ratcliff, G McKoon, Drift Diffusion Decision Model: Theory and data. *Neural Computation* **20**, 873–922 (2008).
10. R Ratcliff, PL Smith, SD Brown, G McKoon, Diffusion decision model: Current issues and history. *Trends in cognitive sciences* **20**, 260–281 (2016).
11. R Bogacz, EJ Wagenmakers, BU Forstmann, S Nieuwenhuis, The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences* **33**, 10–16 (2010).
12. PL Smith, R Ratcliff, Psychology and neurobiology of simple decisions. *Trends in Neurosciences* **27**, 161–168 (2004).
13. R Bogacz, E Brown, J Moehlis, P Holmes, JD Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**, 700–765 (2006).
14. EA Yttri, JT Dudman, Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533**, 1–16 (2016).
15. F Tecuapetla, X Jin, S Lima, R Costa, Complementary Contribution of Striatal Projection Pathways to the Initiation and Execution of Action Sequences. Submitted. *Cell* pp. 1–13 (2016).
16. DM Herz, P Fischer, H Tan, Dynamic control of decision and movement speed in the human basal ganglia Abstract : Intro :. pp. 1–37 (2022).
17. AL de A Marcelino, et al., Pallidal neuromodulation of the explore/exploit trade-off in decision-making. *eLife* **12**, e79642 (2023).
18. CE Geddes, H Li, X Jin, Optogenetic Editing Reveals the Hierarchical Organization of Learned Action Sequences. *Cell* **174**, 32–43.e15 (2018).
19. RL Albin, AB Young, JB Penney, The functional anatomy of disorders of the basal ganglia. *Trends in Neurosciences* **18**, 63–64 (1995).
20. MR DeLong, Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences* **13**, 281–285 (1990).
21. CC Lo, XJ Wang, Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience* **9**, 956–63 (2006).
22. R Bogacz, K Gurney, The Basal Ganglia and Cortex Implement Optimal Decision Making Between Alternative Actions. *Neural Computation* **19**, 442–477 (2007).
23. K Dunovan, T Verstynen, Believer-Skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in Neuroscience* **10**, 1–15 (2016).
24. K Dunovan, C Vich, M Clapp, T Verstynen, J Rubin, Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. *PLoS computational biology* **15**, e1006998 (2019).
25. S Bariselli, WC Fobbs, MC Creed, AV Kravitz, A competitive model for striatal action selection. *Brain research* pp. 0–1 (2018).
26. R Bogacz, EM Moraud, A Abdi, PJ Magill, J Baufreton, Properties of neurons in external globus pallidus can support optimal action selection. *PLoS Computational Biology* **In press**, 1–28 (2016).
27. C Vich, M Clapp, JE Rubin, T Verstynen, Identifying control ensembles for information processing within the cortico-basal ganglia-thalamic circuit. *PLOS Computational Biology* **18**, e1010255 (2022).
28. MJ Frank, Linking across levels of computation in model-based cognitive neuroscience. *An introduction to model-based cognitive neuroscience* pp. 159–177 (2015).
29. C Vich, K Dunovan, T Verstynen, J Rubin, Corticostriatal synaptic weight evolution in a two-alternative forced choice task: a computational study. *Communications in Nonlinear Science and Numerical Simulation* **82**, 105048 (2020).
30. M Clapp, et al., Cbgtpy: An extensible cortico-basal ganglia-thalamic framework for modeling biological decision making. *bioRxiv* (2023).
31. TV Wiecki, I Sofer, MJ Frank, Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics* p. 14 (2013).
32. A Fengler, K Bera, ML Pedersen, MJ Frank, Beyond drift diffusion models: Fitting a broad class of decision and reinforcement learning models with hddm. *Journal of cognitive neuroscience* **34**, 1780–1805 (2022).
33. J Masís, T Chapman, JY Rhee, DD Cox, AM Saxe, Strategically managing learning during perceptual decision making. *eLife* **12**, 1–43 (2023).
34. M Zacksenhouse, R Bogacz, P Holmes, Robust versus optimal strategies for two-alternative forced choice tasks. *Journal of Mathematical Psychology* **54**, 230–246 (2010).
35. O Hikosaka, MK Rand, S Miyachi, K Miyashita, Learning of sequential movements in the monkey: process of learning and retention of memory. *Journal of neurophysiology* **74**, 1652–1661 (1995).
36. F Balci, et al., Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, and Psychophysics* **73**, 640–657 (2011).
37. G Dutilh, J Vandekerckhove, F Tuerlinckx, EJ Wagenmakers, A diffusion model decomposition of the practice effect. *Psychonomic Bulletin and Review* **16**, 1026–1036 (2009).
38. S Uehara, F Mawase, AS Therrien, K Cherry-Allen, P Celnik, Interactions between motor exploration and reinforcement learning. *Journal of Neurophysiology* **122**, 797–808 (2019).
39. EG Liquin, A Gopnik, Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition* **218**, 104940 (2022).
40. T Hanks, R Kiani, MN Shadlen, A neural mechanism of speed-accuracy tradeoff in macaque area lip. *Elife* **3**, e02260 (2014).
41. K Banaie Boroujeni, M Oemisch, SA Hassani, T Womelsdorf, Fast spiking interneuron activity in primate striatum tracks learning of attention cues. *Proceedings of the National Academy of Sciences* **117**, 18049–18058 (2020).
42. A Pasupathy, EK Miller, Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* **433**, 873–876 (2005).
43. JF Cavanagh, et al., Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience* **14**, 1462–1467 (2011).
44. Ka Zaghloul, et al., Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *Journal of Neuroscience* **32**, 2453–2460 (2012).
45. MJ Frank, A Scheres, SJ Sherman, Understanding decision-making deficits in neurological conditions: Insights from models of natural action selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**, 1641–1654 (2007).
46. E Dahlin, L Bäckman, AS Neely, L Nyberg, Training of the executive component of working memory: subcortical areas mediate transfer effects. *Restorative Neurology and Neuroscience* **27**, 405–419 (2009).
47. L Tremblay, JR Hollerman, W Schultz, Modifications of reward expectation-related neuronal activity during learning in primate striatum. *Journal of Neurophysiology* **80**, 964–977 (1998).
48. K Murayama, M Matsumoto, K Izuma, K Matsumoto, Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences* **107**, 20911–20916 (2010).
49. D Shohamy, Learning and motivation in the human striatum. *Current Opinion in Neurobiology* **21**, 408–414 (2011).
50. S Peters, E Crone, Increased striatal activity in adolescence benefits learning. *Nature Communications* **8**, 1983 (2017).
51. MM Yartsev, TD Hanks, AM Yoon, CD Brody, Causal contribution and dynamical encoding in the striatum during evidence accumulation. *eLife* **7**, 1–24 (2018).
52. YR Wu, R Levy, P Ashby, RR Tasker, JO Dostrovsky, Does stimulation of the GPi control dyskinesia by activating inhibitory axons? *Movement Disorders* **16**, 208–216 (2001).
53. T Hashimoto, CM Elder, MS Okun, SK Patrick, JL Vitek, Stimulation of the subthalamic nucleus changes the firing pattern of pallidal neurons. *Journal of Neuroscience* **23**, 1916–1923 (2003).
54. KW McCairn, RS Turner, Deep brain stimulation of the globus pallidus internus in the parkinsonian primate: Local entrainment and suppression of low-frequency oscillations. *Journal of Neurophysiology* **101**, 1941–1960 (2009).
55. W Wei, JE Rubin, Xj Wang, Role of the Indirect Pathway of the Basal Ganglia in Perceptual Decision Making. *Journal of Neuroscience* **35**, 4052–4064 (2015).
56. KN Gurney, MD Humphries, P Redgrave, A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS Biology* **13**, e1002034 (2015).
57. J Baladron, A Nambu, FH Hamker, The subthalamic nucleus-external globus pallidus loop biases exploratory decisions towards known alternatives: a neuro-computational study. *European Journal of Neuroscience* **49**, 754–767 (2019).
58. FA Fortin, FMD Rainville, MA Gardner, M Parizeau, C Gagne, Deap: Evolutionary algorithms made easy. *Journal of Machine Learning Research* **13**, 2171–2175 (2012).

Supporting Information Appendix (SI)

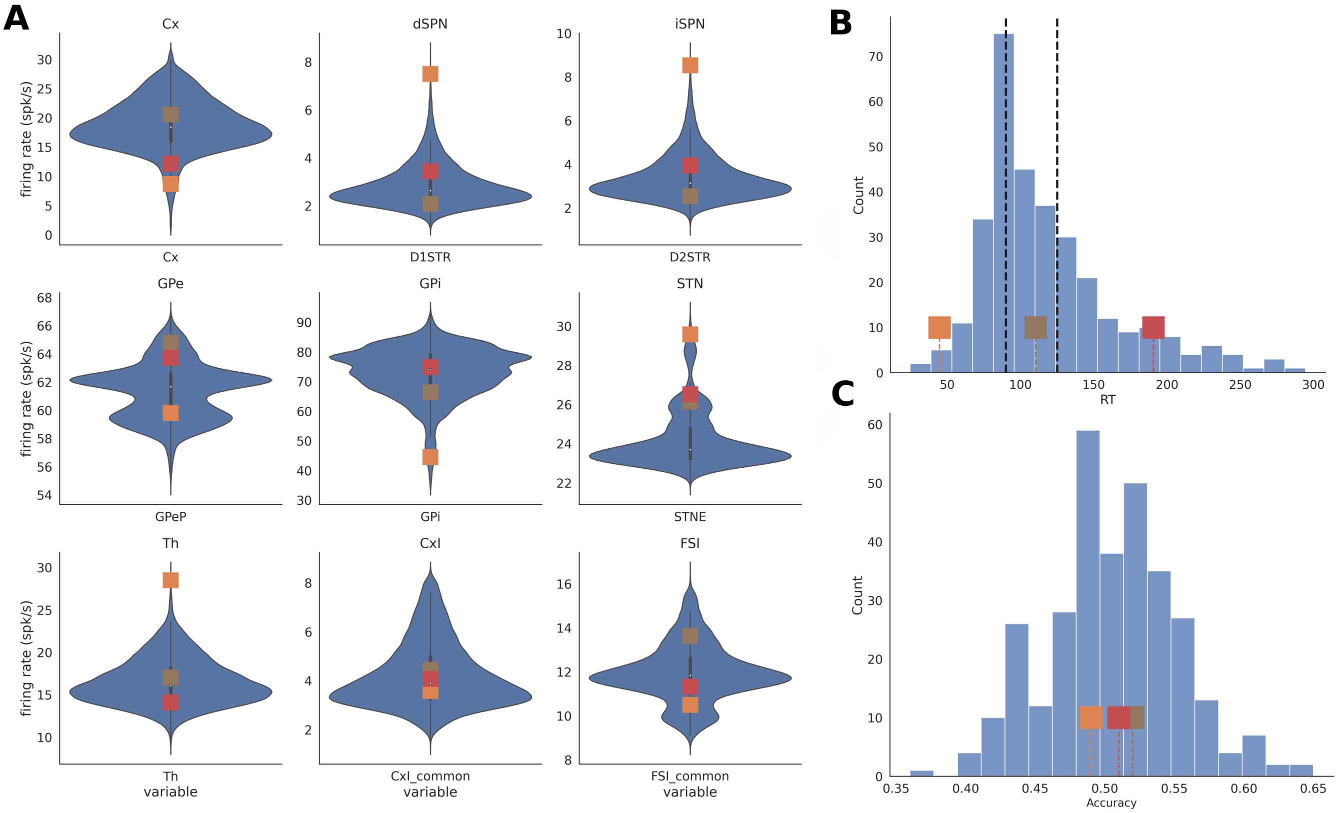

Fig. S1. Network firing rates, accuracy and RTs before plasticity. **(A)** The distributions of average firing rates for the 9 nuclei based on 300 networks. The average firing rates for one example each from three categories of network – fast (orange), intermediate (brown) and slow (red) – are marked on the distribution. The networks before plasticity were categorized as fast, intermediate and slow based on a tertiary split of the reaction time (RT) distribution as shown in **(B)**. The RTs for the exemplar fast (orange), intermediate (brown) and slow (red) networks are marked. **(C)** The average accuracy of all 300 networks. The accuracy is centered around 50% (0.5) because the networks had not undergone plasticity.

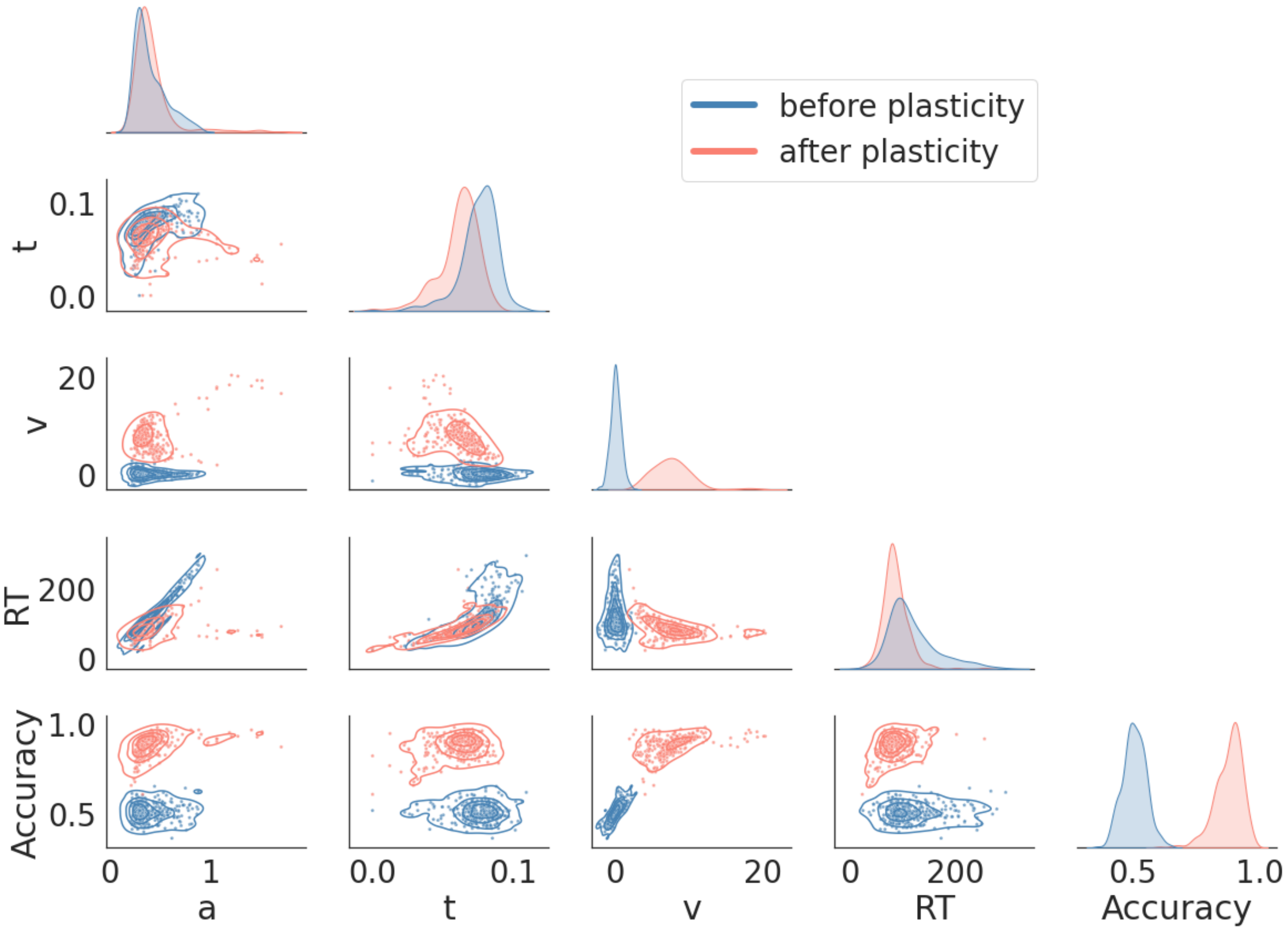

Fig. S2. Comparison of DDM and behavioral measures for all 300 networks before (blue) and after (pink) plasticity. The subplots on the diagonal represent the marginal distributions for DDM parameters $(a,\ t,\ v)$ and behavioral features (RT and accuracy). The onset delay (t) shows a decrease, the drift rate (v) shows an increase, RTs show a decrease, and accuracy shows an increase after plasticity. The off-diagonal subplots show the pairwise covariances.

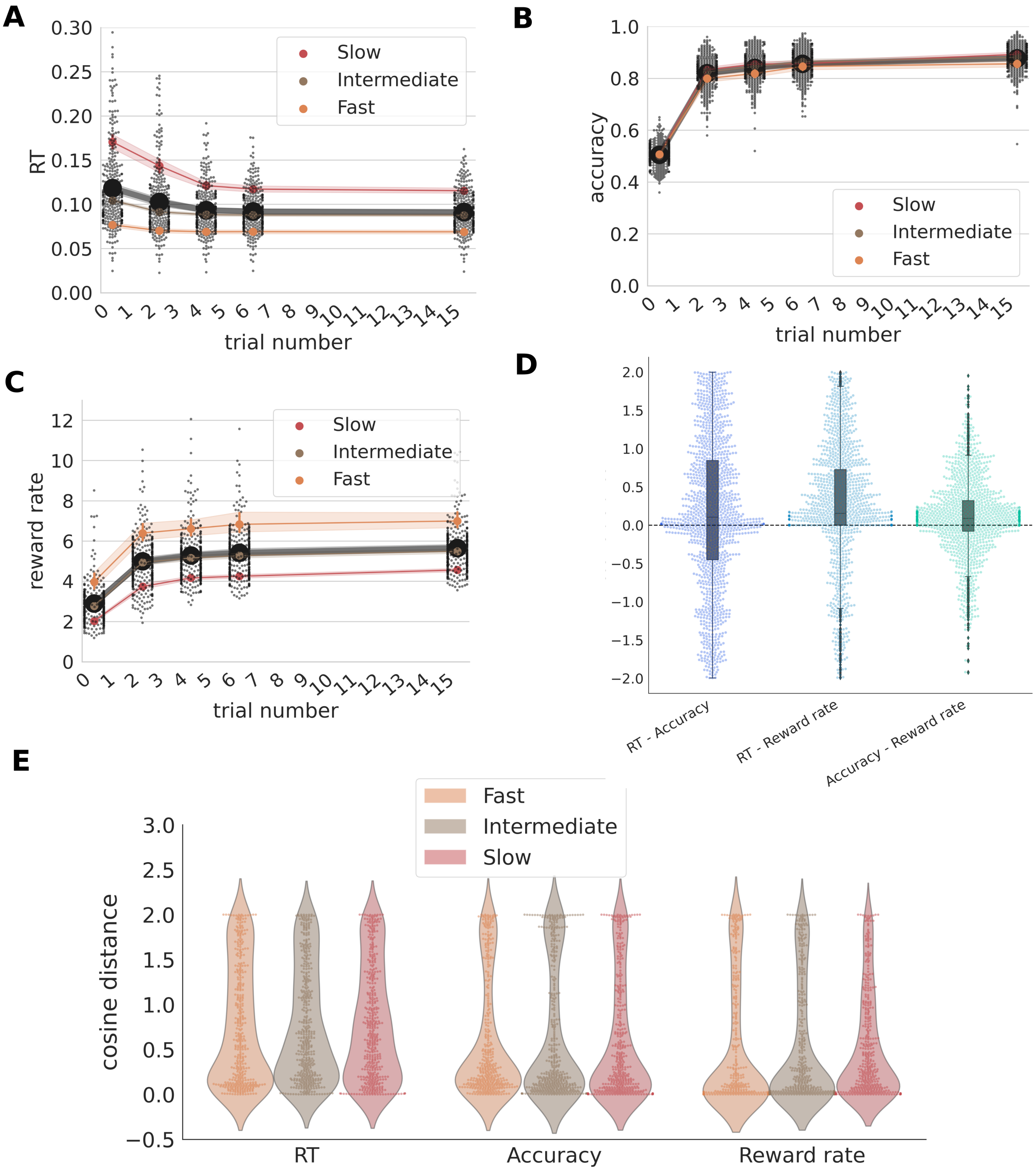

Fig. S3. Evolution of behavioral measures for 300 networks over 16 trials with plasticity. **(A)** Network behavior was assessed after each of 2, 4, 6, 9 and 15 trials. The RTs steadily decreased for all three network categories: fast (orange), intermediate (brown) and slow (red). The average over all 300 networks also showed a steady decrease as shown in black markers and lines. **(B)** The accuracy for the three categories of the networks and the average over all 300 networks increased with plasticity. **(C)** The reward rate for three categories of network and the average over 300 networks increased with plasticity. **(D)** The distribution of differences in cosine distance, measured relative to the direction of greatest increase, for changes in RT vs accuracy, RT vs reward rate, and accuracy vs reward rate for all 300 networks and all stages of plasticity. The comparisons with reward rate yield distributions skewed to significantly above 0, suggesting that the cosine distances are lowest for reward rates. **(E)** Absolute cosine distance distributions shown separately for the three network classes, fast (orange), intermediate (brown) and slow (red).

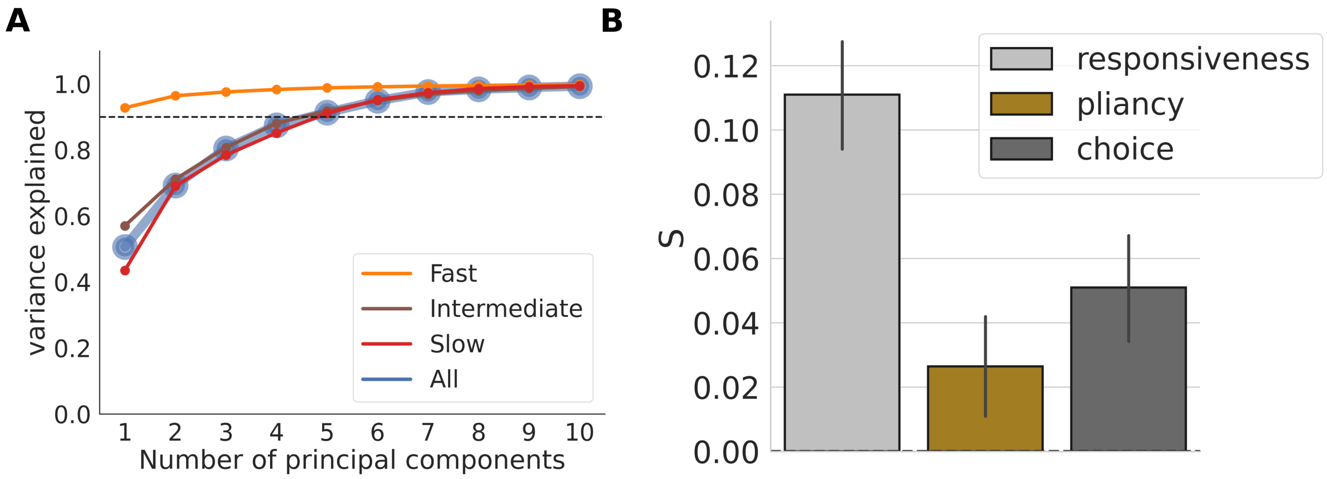

Fig. S4. The least squares solution S pooled over the network types. **(A)** Cumulative variance explained by the first 10 principal components (PC) derived from the changes in firing rates from before to after plasticity. The dashed line indicates 90% of the explained variance. The analysis was done for all the networks pooled together (blue line) and separately for fast (orange), intermediate (brown) and slow (red) networks. For all networks pooled together as well as the separated slow and intermediate networks, the first 5 PCs explain more than 90% of the variance, whereas for fast networks 1 PC suffices. **(B)** The weighted sum of the columns of S (see main text - Fig 4B), pooled over all three network classes (fast, intermediate and slow), shows that the observed changes in firing rates correspond to increased loadings of the responsiveness, pliancy and choice ensembles of the CBGT network.

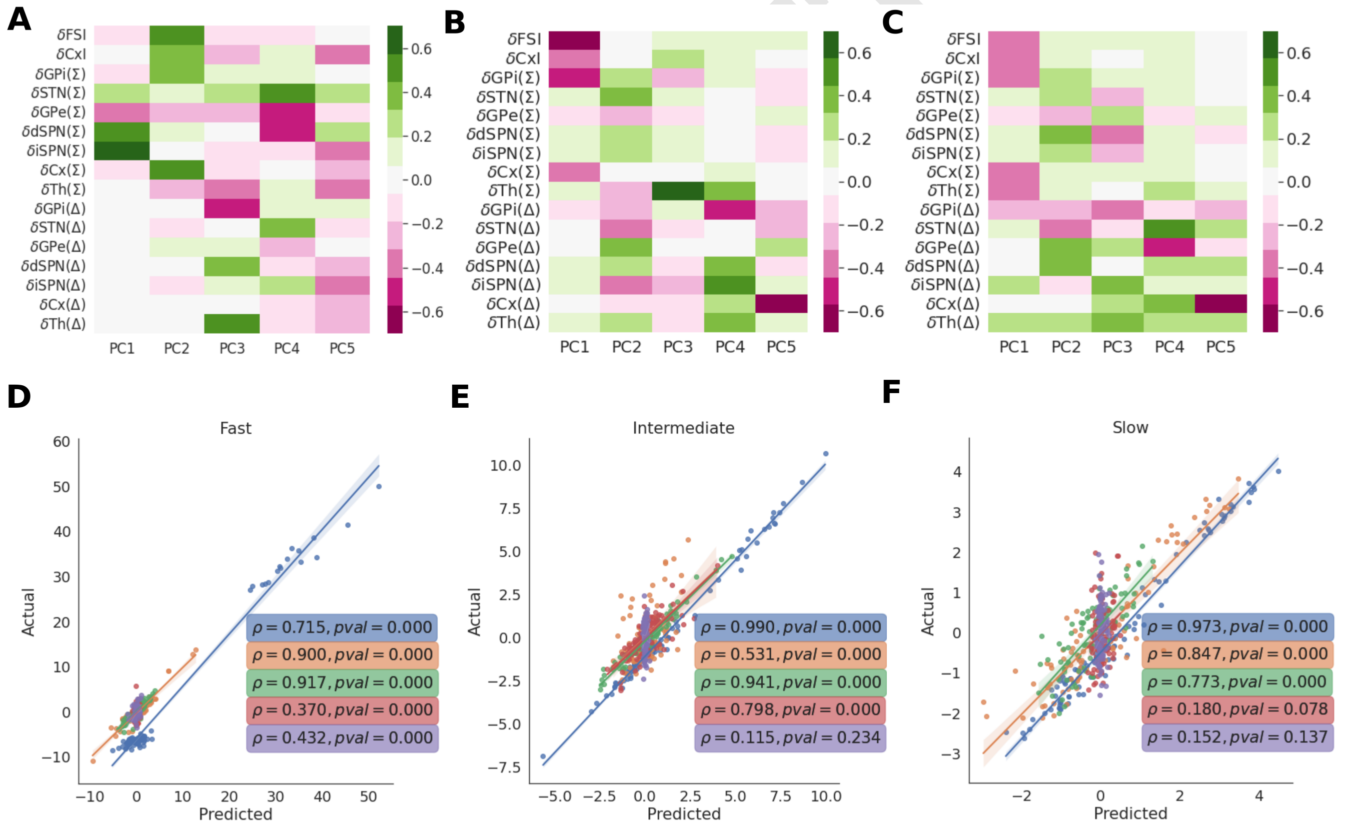

Fig. S5. Reconstruction of firing rate changes from the least squares solution S for the three network classes. **(A)** The first 5 PCs for the firing rate changes in the fast networks. Although the 1st PC explains around 90% of the variance for fast networks, we used 5 PCs to calculate S coefficients (Fig 4C) to be consistent with slow and intermediate networks (Supp Figure S4A). **(B,C)** Same as **(A)** for intermediate and slow networks, respectively. **(D-F)** The dot products of the CCA component vector (C) with each of the 5 columns of S, the least squares solution of $P = CS$, provide an approximate reconstruction of the 5 PCs of the changes in firing rate from before to after plasticity, (ΔF). The quality of the reconstruction was checked by projecting ΔF onto the original PCs for each network (marked as *Actual* on y-axis) and comparing the results with the projections of ΔF onto the reconstructed PCs (marked as *Predicted* on x-axis). The goodness of fit is calculated as the Spearman rank correlation (ρ) between the actual and predicted values. For fast networks **(D)**, the rank correlations (ρ) are high and significant ($p < 0.0001$) for all of the PCs as shown, suggesting that the reconstruction is excellent. For intermediate networks **(E)**, the rank correlations are significant for all PCs except the 5th PC. For slow networks **(E)**, the rank correlations are significant for all except 4th and 5th PCs.

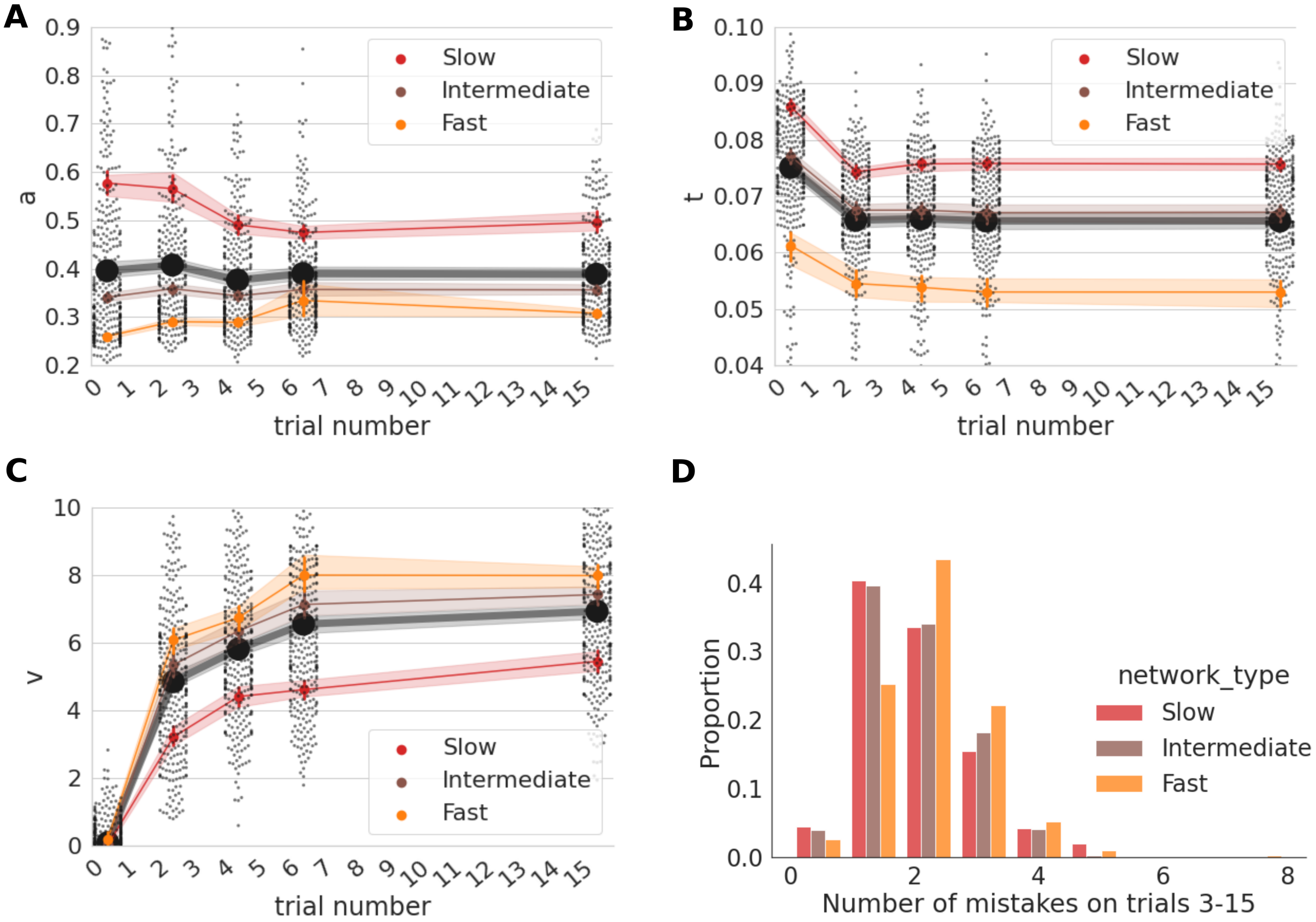

Fig. S6. Evolution of DDM parameters with plasticity. **(A)** The change in boundary height (a) due to plasticity is dependent on network type: slow networks (red) show a decrease, intermediate (brown) show little change, and fast (orange) networks show a slight increase. The mean over all networks is shown by large black circles. **(B)** All network types show a strong decrease in decision onset time (t) due to plasticity. **(C)** All network types show an increase in drift rate (v) due to plasticity. **(D)** Fast networks make more mistakes on average. Shown are the histograms of proportion of unrewarded ("U") trials encountered by all the three network classes after the first two plasticity trials.

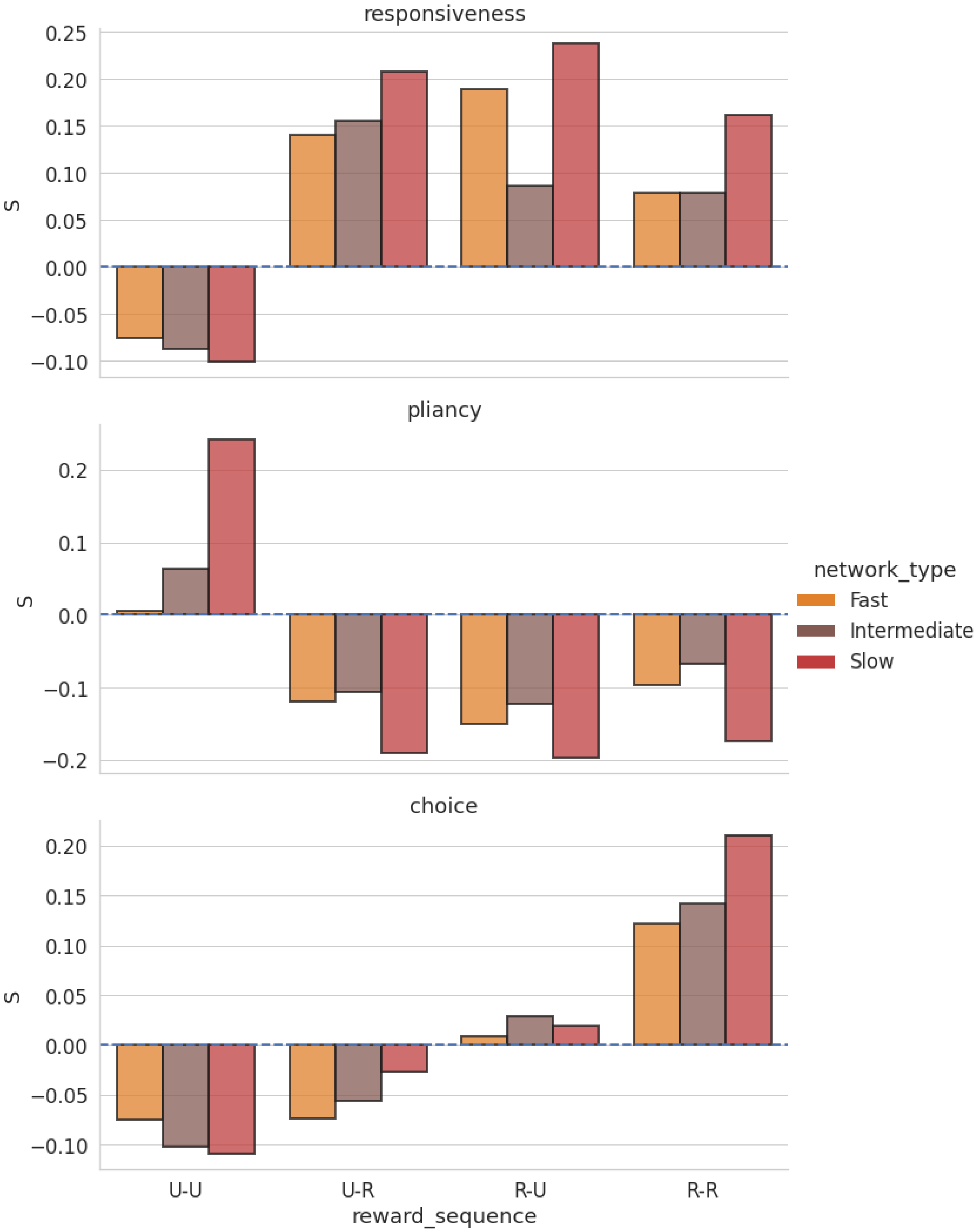

Fig. S7. Effect of reward sequences on the weighting coefficients S for the three network classes. The weighting coefficients S shown in Fig. 5A combine the three network types. The separated coefficients here show the same trends as the combined ones.

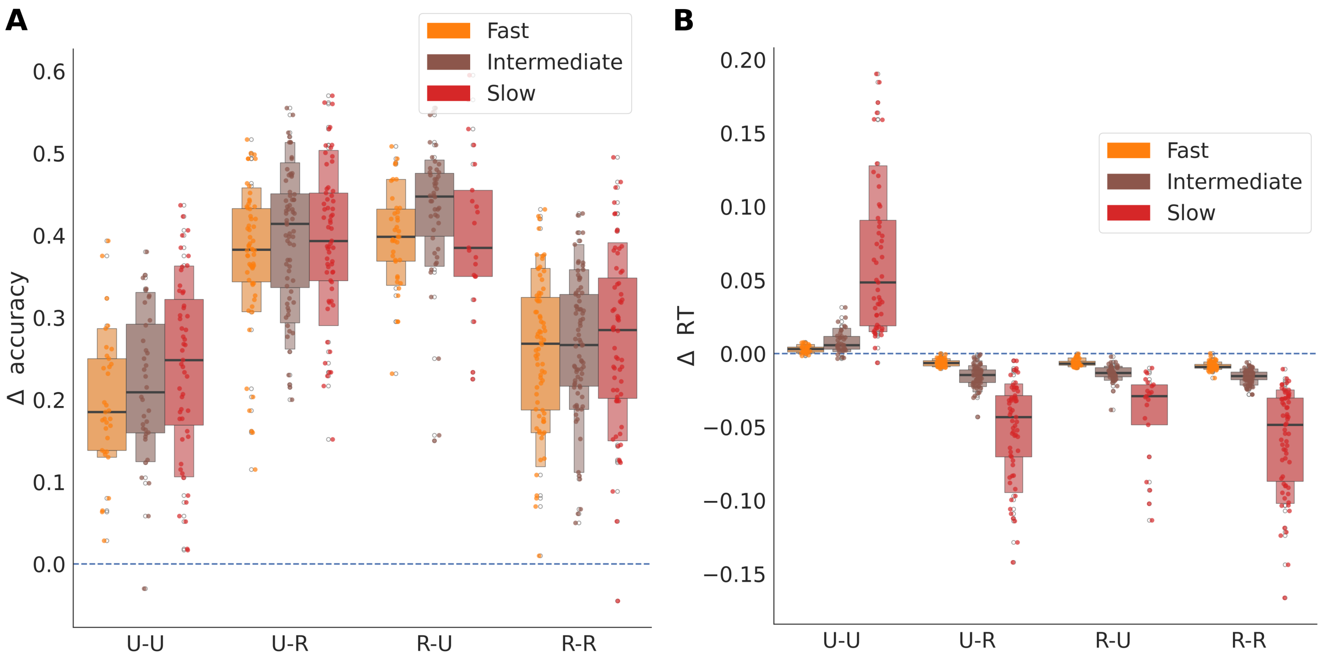

Fig. S8. Effect of reward sequences on changes in accuracy and reaction times (RTs). **(A)** The change in accuracy showed an increase in all cases, but to different extents. The highest increase in accuracy was for one rewarded and one unrewarded trial (U-R and R-U), due to strengthening of the cortico-striatal projection to dSPNs of the optimal choice along with strengthening of cortico-striatal projections to iSPNs of the sub-optimal choice. **(B)** The change in RTs after plasticity for the four outcome sequences. All sequences involving at least one rewarded trial showed a decrease in RT, whereas the sequence with two consecutive unrewarded trials (U-U) showed an increase in RT.

Table S1. Relative number of instances of the reward sequences encountered by each network type. Slow networks encounter a relatively higher proportion of two consecutively unrewarded choices (U-U) as compared to intermediate and fast networks.

Network type	Reward sequence	Relative number of instances (%)
Fast	R-R	36.27%
Fast	R-U	18.62%
Fast	U-R	28.9%
Fast	U-U	16.2%
Intermediate	R-R	35.1%
Intermediate	R-U	19.9%
Intermediate	U-R	29.4%
Intermediate	U-U	15.6%
Slow	R-R	32.1%
Slow	R-U	10.7%
Slow	U-R	31.1%
Slow	U-U	26.0%