Collicular circuits for flexible sensorimotor routing

Chunyu A. Duan[1,2,3]*, Marino Pagan[2,3]*, Alex T. Piet[2,3], Charles D. Kopec[2,3], Athena Akrami[2,3,4], Alexander J. Riordan[2,3], Jeffrey C. Erlich[2,3,5] & Carlos D. Brody[2,3,4#]

[1] Institute of Neuroscience, State Key Laboratory of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China.
[2] Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08544, USA.
[3] Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA.
[4] Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey 08544, USA.
[5] NYU-ECNU Institute of Brain and Cognitive Science, NYU Shanghai, Shanghai, China.

*These authors contributed equally to this work.
[#]Correspondence should be addressed to C.D.B. (brody@princeton.edu)

ABSTRACT

Flexible and fast sensorimotor routing, based on relevant environmental context, is a central component of executive control, with prefrontal cortex (PFC) thought of as playing a critical role and the midbrain superior colliculus (SC) more traditionally viewed as the output of cortical flexible routing. Here, using a rat task in which subjects switch rapidly between task contexts that demand changes in sensorimotor mappings, we report that silencing of the SC during a delay period, during which task context is encoded in SC activity, impaired choice accuracy. But inactivations during the subsequent choice period, during which the subject selects their motor response, did not. Furthermore, a defined subset of SC neurons encoded task context more strongly than PFC neurons, and encoded the subject's motor output choice faster than PFC neurons or other SC neurons. These data suggest cognitive and decision-making roles for the SC. We used computational methods to identify different SC circuit architectures that could account for these results. We found numerous, highly varied SC model circuits that matched our experimental data, including circuits without inhibitory connections between units representing opposite decision outputs. But all successful model circuits had inhibitory connections between units on the same side of the brain representing opposite contexts. This anatomical feature appears to be a key experimental prediction for models in which the SC plays a decision-making role during executive control.

INTRODUCTION

Many biological circuits are complex enough that even high-throughput experiments provide data that constrain only a fraction of the characteristics that fully define the circuit. An often-used computational approach to this problem is to build complete models in which values of the unknown characteristics are chosen or fit such that the model matches its postulated function and known experimental data. These chosen values then constitute hypotheses, or predictions that can guide further experiments[1]. However, even simple circuits can have many different configurations that produce the same data, referred to as having many different "solutions"[2–5] or "sloppiness"[6]. For circuits with many parameters, exhaustive searches for configurations that are also solutions, the approach taken in some cases[2–4], is impossible, because the number of possible configurations grows exponentially with the number of parameters.

We were confronted with this problem when exploring a 12-parameter circuit model of the SC to account for experiments that we report below. These experiments suggested an unexpected cognitive role for the SC during a rat executive control task. To overcome this problem of high dimensionality, we reconfigured the search for circuit solutions as a minimization problem, and used modern algorithmic differentiation methods, which render almost effortless the computation of first- and second-order derivatives of arbitrarily complex models[7]. This approach allowed us to find a wide variety of SC models that were compatible with our data. Comparing across varied solutions allows identifying features that are common to all the solutions, which then constitute key predictions for the entire class of models in the parameter space being explored. In the case of this study of the SC, we expected that inhibitory connections between units representing opposite decision outputs would be necessary[8,9]. But this was not the case: many solutions did not have that feature. In contrast, all solutions found had inhibitory connections between units encoding opposite task contexts on each side of the brain, suggesting that anatomical characteristic as an unexpected but key experimental prediction.

RESULTS

A subset of task-encoding neurons linked to decisions

We used a recently-developed behavior that demonstrates that rats can exert executive control to perform rapid task switching[10]. On each trial of this behavior, rats are first presented with an auditory cue indicating a task context in effect for the current trial (labelled 'Pro' or 'Anti'), followed by a short memory delay period, and finally a choice period during which a visual stimulus to one side is turned on and rats are required to either orient toward ('Pro') or away ('Anti') from it (**Fig. 1a,b**). The choice period is when task context information must be combined with the sensory stimulus in order to produce the correct sensorimotor transformation. Rats can flexibly switch between these two task contexts from one trial to the next[10], and display multiple behavioral asymmetries between Pro and Anti responses (Extended Data Fig. 1), similar to those observed in the primate pro/antisaccade paradigm[11–13]. These asymmetries indicate the Pro task as a stimulus-driven, prepotent task, while the Anti task is more cognitively demanding[10]. Pharmacological inactivations of the SC and the PFC impair rats' ability to perform the Anti task but not the easier prepotent Pro task[10].

Here, to investigate neural representations in SC and PFC, we recorded well-isolated single units in the intermediate and deep layers of the SC (193 neurons; Methods and Extended Data Fig. 2), and in the prelimbic region of the medial PFC (331 neurons; Extended Data Fig, 3d-f), from 7 rats performing the ProAnti task-switching behavior. Individual neurons in the SC and PFC displayed firing rates that depended on, and therefore encoded, task context (Pro or Anti) as well as the subsequent motor choice (Left or Right; **Fig. 1c**). Similar to observations in prefrontal cortex during cognitive tasks[14–16], the encoding in SC appeared to be highly heterogeneous across different neurons, and multiplexed across the whole trial (**Fig. 1d** and Extended Data Fig. 3; [17–19]). To evaluate the amount of task context information in SC versus PFC populations we used a cross-validated linear decoding approach (Methods; [20]). Although both populations contained above-chance task information throughout the trial duration, decoding for whether a trial was Pro or Anti was significantly more accurate in the SC than in the PFC for equally-sized populations ($p<0.01$, **Fig. 1e**, left; n=193), even after controlling for firing rate differences between the two areas (Extended Data Fig. 4a). The stronger task information in the SC could only be matched when the number of neurons in a simulated PFC population was 5 times larger than the SC population (Extended Data Fig. 4b).

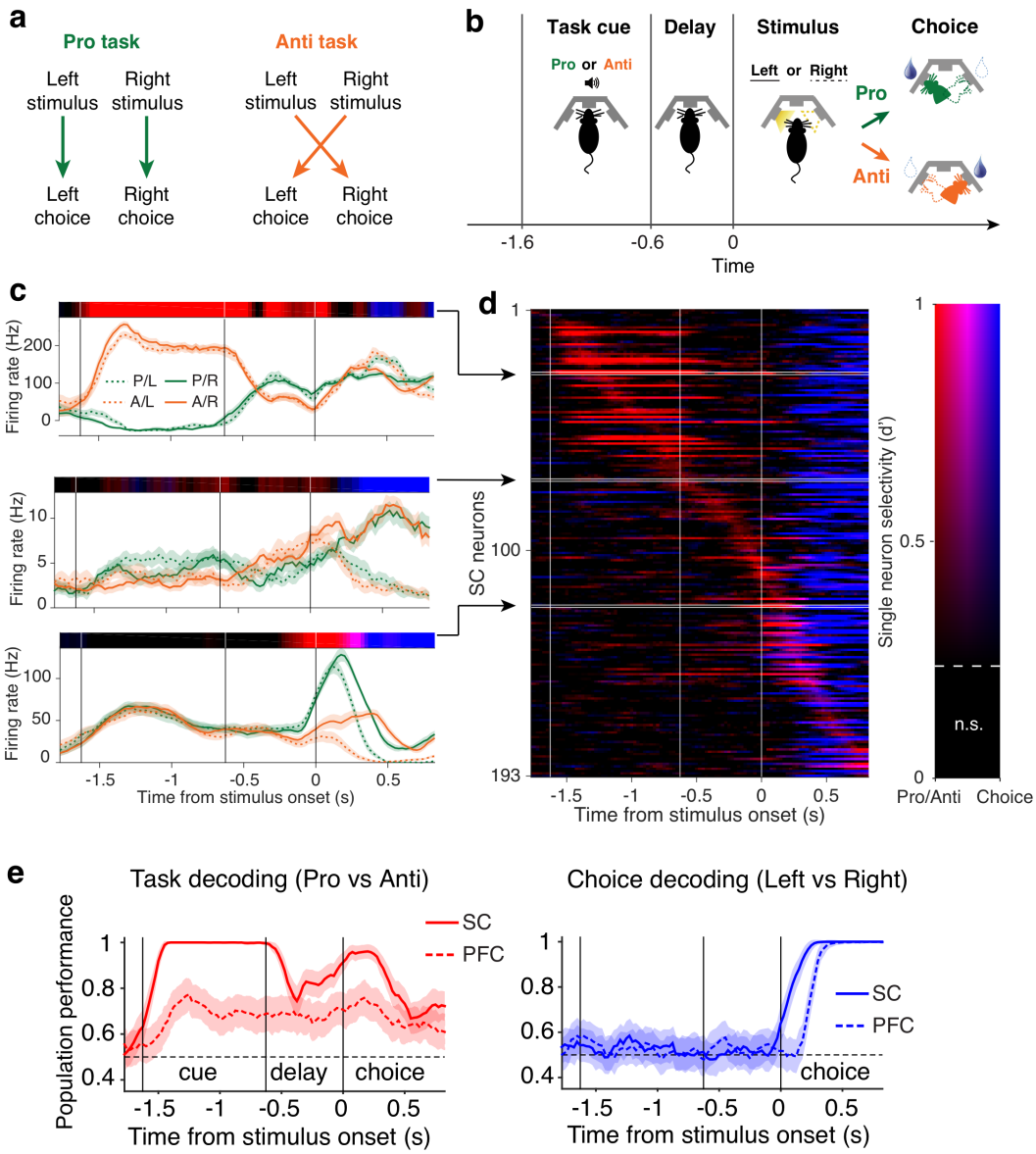


Figure 1 | SC and PFC populations contain task and choice information during rapid sensorimotor task switching. **a**, Rules for the Pro and Anti task contexts. In the Pro task, rats should orient *toward* a lateralized stimulus (left or right) for reward; in the Anti task, rats should orient *away* from the stimulus for reward. Trained rats can switch between these two known task contexts from one trial to the next. **b**, Rats nose poke in the center port to initiate each trial and keep fixation during the task cue (Pro or Anti sound) and delay periods. After the delay, the animal is allowed to withdraw from the center port, and a lateralized light (left or right) is turned on to indicate the stimulus location. Rats then poke into one of the side pokes for reward. **c**, Peri-stimulus time histogram (PSTH) for 3 example SC neurons on Pro-Go-Right (green solid), Pro-Go-Left (green dashed), Anti-Go-Right (orange solid), and Anti-Go-Left (orange dashed) trials. PSTHs are aligned to stimulus onset. Top, task (red) and choice (blue) selectivity as a function of time for each neuron. **d**, Information encoding matrix of the SC population. Each row of the matrix represents the d' of a single neuron as a function of time. The intensity of the color represents how "informative" a neuron is, and the RGB values are associated with different types of information (Pro/Anti, red; choice, blue; mixed, purple). Neurons are sorted by the timing of their peak Pro/Anti d'. d' that are not significant (n.s.) are set to 0. **e**, Evolution of classification performance over time in the SC (solid) and PFC (dashed) population. Left, mean and s.e.m. performance for linear classification of correct Pro versus Anti trials. Spikes are aligned to stimulus onset, and counted over windows of 250 ms with 25-ms shifts between neighboring windows. Note that performance is plotted over the right edge of the window (causal). Right, classification performance to linearly separate Go-Left versus Go-Right trials, similar to the left panel.

In addition, left versus right choice information appeared significantly earlier in the SC than in the PFC (latency difference = 191 ± 23 ms; $p<0.01$, **Fig. 1e**, right). This choice information latency difference is not a result of firing rate differences between the two populations (Extended Data Fig. 4a), and cannot be reduced by increasing the number of PFC neurons in a pseudo-population (Extended Data Fig. 4c). These results argue against a model in which the decision is first computed in PFC and then relayed to SC. Instead, they suggest a critical role of SC across all stages of the ProAnti behavior.

A closer examination of SC neurons revealed subpopulations with distinct activity patterns (**Fig. 2**). For each SC neuron, we computed the temporal profile of significant task selectivity (Pro vs Anti), and ranked neurons by the time of their peak selectivity (**Fig. 2a**). One group of SC neurons (which we labelled "cue neurons", cyan) differentiated between Pro and Anti trials most strongly during the auditory cue, whereas another subpopulation maintained task selectivity most strongly when the auditory cue was no longer present ("delay/choice neurons", yellow). The representation of task context by cue neurons was progressively weakened after the end of the cue, and it did not differentiate between correct and error trials (**Fig. 2b, top**), consistent with a purely sensory signal with little direct relationship to behavior.

In contrast, three lines of evidence suggest that the delay/choice neurons play a key role in behavior. First, their task information slowly ramped up throughout the cue presentation and the delay to peak at the time when rats were required to make a motor choice (**Fig. 2b**, **bottom**, solid line). Second, this representation was entirely disrupted on error trials (**Fig. 2b**, **bottom**, dashed line), indicating a strong correlation with behavior. Third, these neurons contained a very early representation of the correct choice, significantly faster than the SC cue neurons or the PFC neurons (**Fig. 2c**, delay-choice neurons = 84 ± 19 ms, cue neurons = 196 ± 38 ms, PFC neurons = 290 ± 34 ms, $p<0.01$). Thus, the SC delay/choice neurons contain a performance-dependent task context signal (Pro vs Anti) that could be used for correct routing of the upcoming target stimulus information (Left vs Right side light), so as to produce the correct context-dependent orienting choice.

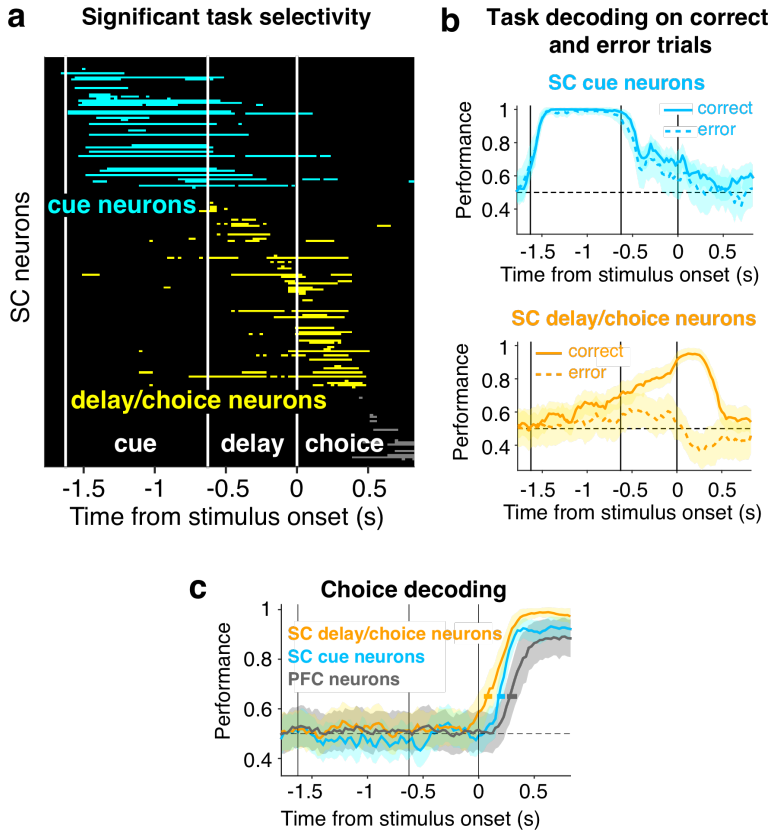


Figure 2 | Distinct roles of SC subpopulations. **a**, Timing of significant Pro/Anti selectivity (d') for all SC neurons, sorted by peak d'. Significance threshold was determined by shuffled data. We separated SC neurons into two groups based on the timing of their Pro/Anti selectivity. "Cue neurons" (cyan, n=29) differentiated between Pro and Anti trials most strongly during the auditory cue; "delay/choice neurons" maintained task selectivity most strongly when the auditory cue was no longer present (yellow, n=45). **b**, Performance of task decoding on correct versus error trials. Linear classifiers trained on correct trials were tested for separate correct trials (solid) or error trials (dashed). The representation of task context by delay/choice neurons was disrupted on error trials whereas such information in the cue neurons did not differentiate between correct and error trials. **c,** Choice decoding performance of SC subpopulations and PFC neurons (n=29 to match number of cue neurons, see Methods). Choice information emerged first in SC delay/choice neurons. Shaded areas (vertical error bars) indicate s.e.m. of decoding accuracy for each population across time. Horizontal error bars represent s.e.m. of the timing of reaching 0.65 decoding accuracy for each population.

We therefore focused on the SC delay/choice neurons to examine how task and choice signals were multiplexed (**Fig. 3**). In our behavior, the sensorimotor transformation occurs immediately after the visual target onset, when animals apply the non-spatial task context (Pro or Anti) to guide spatial orienting responses (which we will describe as either ipsi- or contralateral to the recorded side). In contrast to the heterogeneity initially observed in the entire SC population (**Fig. 1d** and Extended Data Fig. 3), focusing on the subset of delay/choice neurons during this critical time window revealed a systematic relationship between each neuron's task selectivity and choice selectivity (**Fig. 3** and Extended Data Fig. 5). Most neurons that fired more in trials with contralateral orienting responses also fired more during Pro task trials (**Fig. 3a**). Conversely, most Ipsi-preferring neurons were also Anti-preferring (**Fig. 3b**). This suggests the existence of two broad groups of neurons during the choice period. We refer to these two groups of neurons as Pro/Contra-preferring neurons and Anti/Ipsi-preferring neurons.

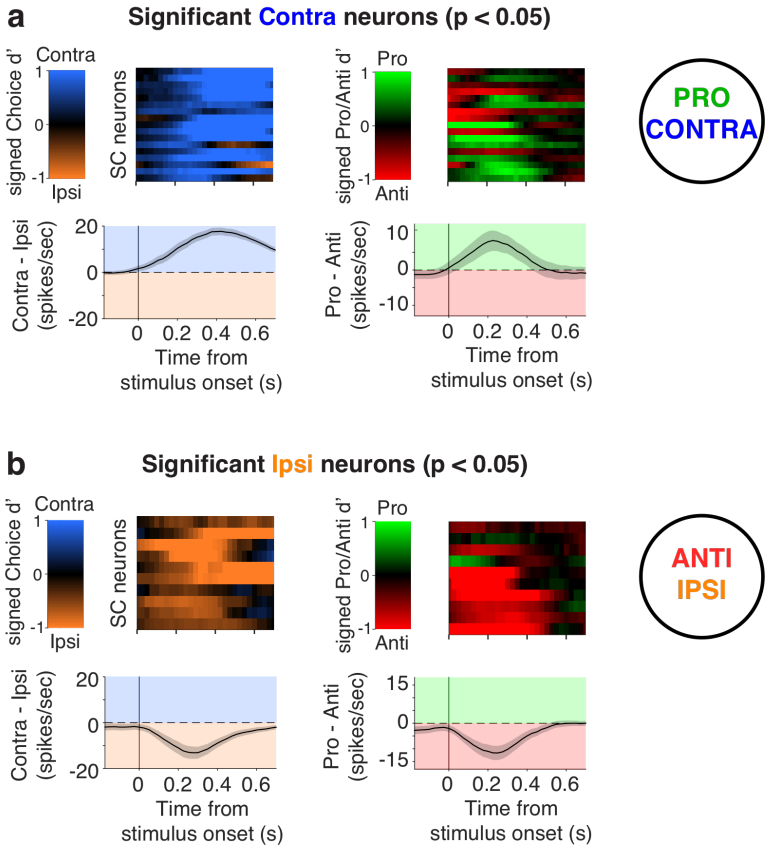


Figure 3 | A relationship between task and choice encoding around stimulus onset, suggesting two groups of neurons. a, Neurons selected as having a significantly greater firing rate on trials when the animal's choice is to orient Contralaterally (n = 17). Left top shows one neuron per row, with the color indicating the strength of firing rate difference between Contra- and Ipsi-orienting trials, as a function time relative to the visual stimulus onset. Left bottom shows firing rate difference (Contra-Ipsi) averaged over these neurons. Right panels are the same neurons as in the left panels, with each row corresponding to the same neuron as the one on the left, but now analyzed for Pro - Anti selectivity and firing rate. Contra-preferring neurons tend to be Pro-preferring neurons. **b,** as in panel a, but showing significantly Ipsi-preferring neurons (n = 10). These tend to also be Anti-preferring.

SC activity is necessary during the task-encoding delay period.

Since different behavioral epochs of the task require very different computations, we selectively probed the requirement of SC activity during separate epochs[21–23] using bilateral optogenetic inactivation of SC neurons, mediated by virally-expressed eNpHR3.0, a light-activated chloride pump (**Fig. 4a, b** and Extended Data Fig. 2c, Methods).

Optogenetic inactivation that covered the entire trial period (3 s) of a randomly selected 25% of trials resulted in a selective Anti impairment in those trials (**Fig. 4c** and Extended Data Fig. 6a; permutation test $p<10^{-3}$ across animals or across all trials). This replicates previous pharmacological inactivation results where SC activity was suppressed during the entire session[10]. Turning to temporally-specific inactivations, we found that bilateral SC inactivation during the task cue period did not result in any behavioral deficit (**Fig. 4d**, left), consistent with a sensory representation role for cue neurons not required for correct performance (**Fig. 2b**). In contrast, bilateral SC inactivation during the delay epoch significantly increased error rates on Anti trials (**Fig. 4d**, middle; bootstrapped $p<0.001$), demonstrating for the first time, that SC delay-period activity is required for maintaining non-spatial information, here a task context. Finally, given the strong and early choice signal in the SC, we were surprised to find that bilateral choice period inactivation did not have any effect on choice accuracy (**Fig. 4d**,

right; p>0.05), although we did observe that correct Anti responses were slightly slowed down after choice period SC inactivation (22.5 ± 15.3 ms, p<0.05; Extended Data Fig. 6b).

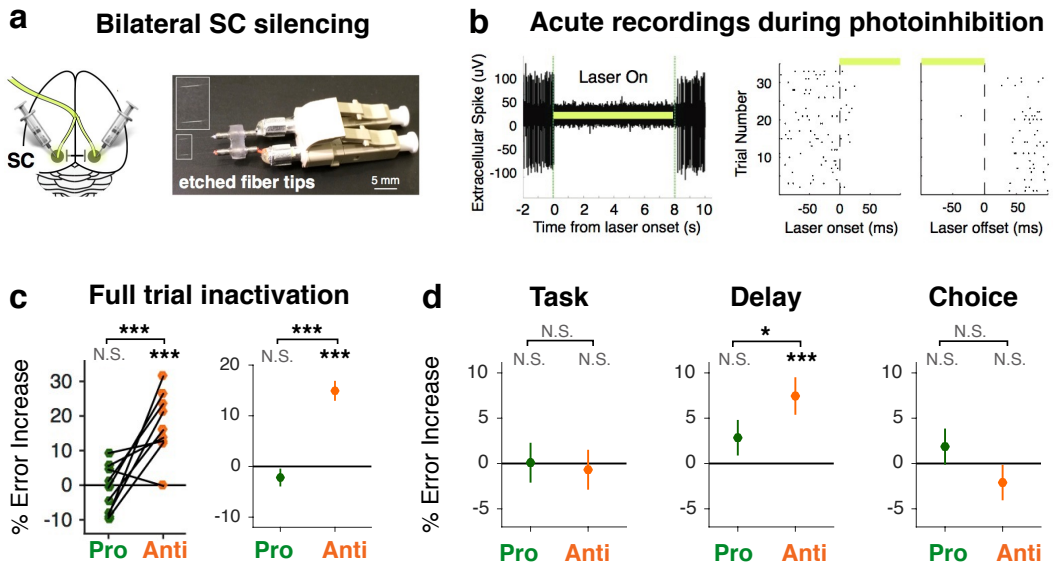


Figure 4 | SC delay activity is required for the Anti task. a, Experimental design for bilateral SC optogenetic inactivation. Left, a schematic that indicates virus infection and laser stimulation in the SC on both hemispheres. Right, an example of the optical fiber implant. The taper of each fiber is chemically sharpened to be approximately 2 mm long for stronger and more unified light delivery. The distance between the two fibers are constructed to be exactly 3.6 mm to target bilateral SC. **b**, Physiological confirmation of optogenetic inactivation effect in an anesthetized animal. Left: acute extracellular recording of spontaneous activity in the SC expressing eNpHR3.0. Laser illumination period (8 s) is marked by the light green bar. Right: spike activity aligned to laser onset and laser offset over multiple trials. Note that the onset and offset of the inhibitory effect are on the scale of tens of milliseconds. **c**, Effect of full-trial inactivation of bilateral SC. Mean Pro (green) and Anti (orange) error rate increase due to SC inactivation for all individual rats (n=9, left) and across all trials (Pro=662 trials, Anti=615 trials, right). Left, each data point represents the mean effect across sessions for a single rat. Right, means and s.e.m. across trials (concatenated across all 60 sessions). **d**, Effect of sub-trial inactivations of bilateral SC on Pro and Anti error rate (mean and s.e.m. across trials from 102 sessions). Statistical comparison between Pro and Anti effects were computed using a permutation test, shuffled 5000 times. N.S. p>0.05; *p < 0.05; ***p < 0.001. Note that all types of inactivations were randomly interleaved for each session.

Many model collicular circuits consistent with the experimental data.

Could neural circuitry between Pro/Contra and Anti/Ipsi neurons within the SC lead to the pattern of results seen in our optogenetic experiments (**Fig. 4d**)? Or would choice formation circuitry external to the SC be necessary to explain the lack of an inactivation effect during the choice period (**Fig. 4d,** right)? To address these questions, we explored computational models in which the SC was represented by two groups of neurons, Pro/Contra and Anti/Ipsi neurons, on each side of the brain (**Fig. 5a**). Since unilateral SC stimulation drives contralateral orienting motions[24–26], we took the Pro/Contra neurons as driving the motor output, with the final choice determined by which of the two Pro/Contra units had the greater activity[27]. The model had free parameters describing the sign and strength of connections between the units (**Fig. 5a**), magnitude of a noise parameter, degree of silencing induced by optogenetic inactivation, and others (Methods), for a total of 12 free parameters. Connections between the two sides of the SC can be both excitatory and inhibitory[28,29], so we made no assumptions as to connection signs. A set of parameter values that successfully reproduced the experimental data would constitute a hypothesis regarding SC circuitry, and its connectivity parameters would constitute anatomical predictions that followed from that hypothesis. Guided by our intuitions, we found one such set of parameter values. But we quickly found that our intuitions provided only a limited understanding of the range of dynamics possible, even for this simplified 4-dimensional model, and intuition alone was not sufficient to answer whether the predictions that followed from this set of parameters were necessary or incidental. We therefore turned to automated methods for a more complete search of the space of solutions. We wrote down a scalar cost function of

the parameters, J, such that J would be low if the following conditions were met and would be high otherwise (Methods):

1. Choice period inactivation had no effect on choice accuracy (**Fig. 4d**, right).
2. Delay period inactivation impaired accuracy on Anti trials but not Pro trials (**Fig. 4d**, middle).
3. During control trials (no inactivation) the fraction of correct Pro trials was higher than on Anti trials[10].

We then optimized J starting from many different random parameter values. The optimizations that produced low values of J had final parameter values that we refer to as solutions. 18,500 initial random seeds, none of which were themselves solutions, produced 354 solutions. We clustered the dynamics (activity as a function of time for the four units in the model) produced by those 354 solutions into 7 groups (Methods), and found the 2D linear projections of the dynamics that best separated those clusters (**Fig. 5b**). The solutions represented disjoint sets of a wide variety of very different dynamics and parameter values (Extended Data Fig. 7). We conclude that collicular circuitry involving Pro/Contra and Anti/Ipsi neurons is sufficient, in a broad variety of configurations, to reproduce a lack of a choice effect during choice period silencing but impairment of Anti trials during delay period silencing.

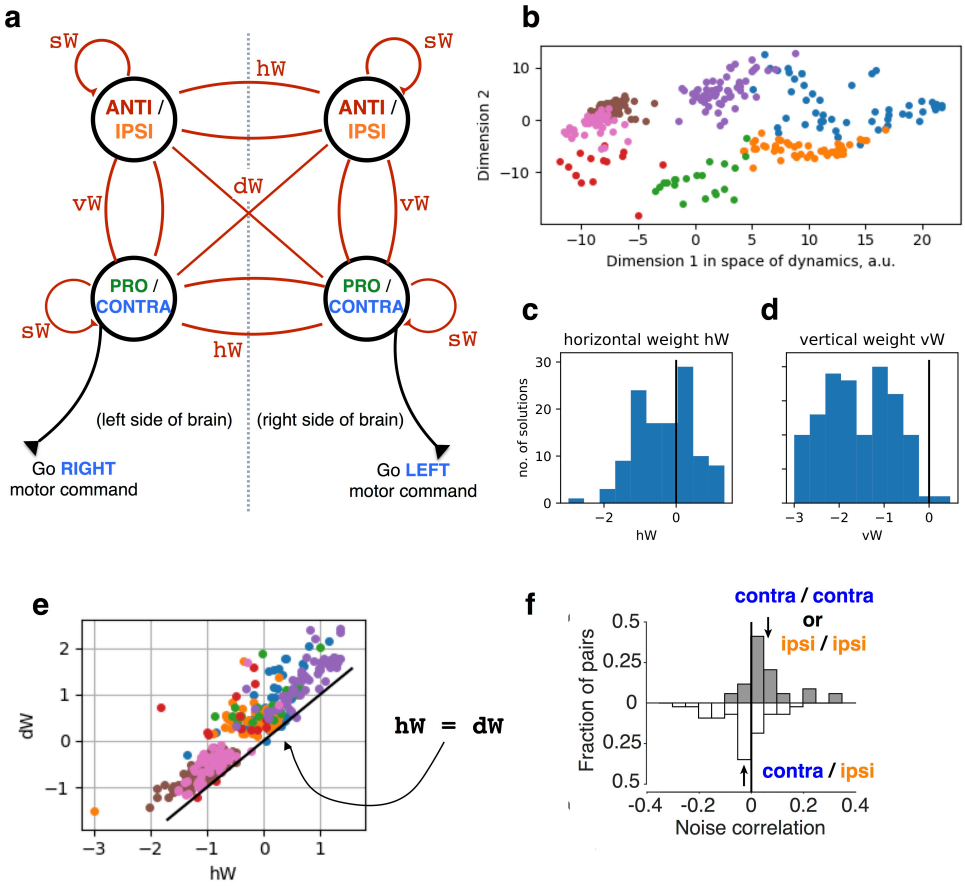


Figure 5 | Model and experimental data consistent with its predictions. a. Schematic of SC model, showing four units, each representing a population of SC delay/choice neurons. All connections are bidirectional. **b.** Projection of model solutions onto the 2D space that maximally explains variance in model dynamics across the set of model solutions. Individual dots are unique model solutions, color coding reflects the output of clustering on the model solutions. **c.** Histogram of horizontal weights in model solutions, showing a wide variety of values of mixed sign. **d.** Histogram of vertical weights in model solutions, showing a wide variety, but almost entirely negative. **e.** Scatter plot of horizontal weights against diagonal weight. Equality line added for reference. **f.** Trial-by-trial noise correlation between pairs of simultaneously recorded neurons on one side of the SC, calculated for within-group pairs of neurons (Contra/Contra or Ipsi/Ipsi, upper histogram) and between-group pairs (Contra/Ipsi). Noise correlation distribution for within-group pairs was significantly shifted above 0 (mean = 0.082 ± 0.02; $p<0.01$), whereas the between-group distribution was significantly shifted below 0 (mean = -0.052 ± 0.02; $p<0.05$), consistent with negative vW as predicted by the model. Arrows indicate the mean values.

Decision-making neural dynamics are often postulated as being driven by mutual inhibition[8,9]. We therefore expected inhibitory connections between the two Pro/Contra units, whose mutual competition determines the decision output, as a necessary feature in all solutions (hW<0 in **Fig. 5a**). However, this intuition proved to be incorrect (**Fig. 5c**). In contrast, two features were observed across all solutions. First, the connection between opposite task units on opposite sides of the brain (dW) was always slightly more positive than the connection between units representing the same task but opposite choices (hW), even across disjoint sets of solutions with widely varying values of the individual parameters hW and dW (**Fig. 5e**). Second, all but one solution found had inhibitory connections between units representing opposite tasks on the same side of the brain (vW<0; **Fig. 5d**), and that sole solution had a very weak connection. We therefore predict that there should be inhibitory connections between Anti/Ipsi and Pro/Contra neurons on the same side of the brain. This prediction is consistent with negative noise correlations between those pairs of neurons that we observed experimentally (**Fig. 5f**).

The computational searches thus proved critical to identifying reliable experimental predictions. Initial predictions guided by intuition were shown to be irrelevant (**Fig. 5c**) and other, unexpected predictions were shown to be robust (**Fig. 5d,e**).

DISCUSSION

Understanding the neural mechanisms of cognition provides a critical foundation for better assessment and potential treatment of psychiatric disorders. Motivated by seminal studies where patients with prefrontal cortex (PFC) lesions[30,31] or schizophrenia[32,33] failed to perform the antisaccade task, decades of primate research have focused on the PFC as the key neural substrate underlying executive functions, especially inhibitory control of downstream motor areas[11,34,35]. Meanwhile, evidence against the cortical inhibition model[36,37] or evidence for an inhibitory role of the human superior colliculus (SC) during antisaccades[38] were largely overlooked until recently[39,40].

Our analysis of prefrontal and midbrain activity during flexible sensorimotor routing provides three lines of evidence for an extended executive control network that includes the SC. First, SC neurons contain stronger task context information and earlier decision information than PFC neurons (**Figs. 1,2**). Second, although agnostic about the origin of task context representation, we show that the task context information within the SC is causally required for context-based behavioral flexibility (**Fig. 4**). Finally, our computational models demonstrate that fast sensorimotor routing can be achieved through control of nonlinear dynamics within a collicular circuit (**Figs. 3,5**). As proposed circuit models in biology grow in complexity, solution degeneracy overtakes human intuition[5,6]. Using computational approaches to search the space of solutions becomes necessary for identifying reliable model predictions and thus critical if models are to usefully guide experiments (**Fig. 5**). By examining common features across distributions of successful yet varied solutions, we identified inhibition between competing task context representations to be an essential component of a collicular circuit model of executive control. Together, our experimental and modeling work suggest that cognitive functions that are normally associated with PFC circuit[14,15,41] could be equally attributed to the midbrain SC[42,43]. These results call for a broadening of focus in basic and clinical studies of executive functions to include interconnected cortical and subcortical areas[39,44,45].

METHODS

Subjects. Nineteen adult male Long-Evans rats (Taconic) were used for the experiments presented in this study. Of these, 7 rats were used for electrophysiology recordings, and 12 rats were implanted with optical fibers for the optogenetic inactivation and YFP control experiments. Animal use procedures were approved by the Princeton University Institutional Animal Care and Use Committee and carried out in accordance with NIH standards.

Behavior. Rats were trained on the ProAnti task-switching behavior[10]. Each trial began with an LED turning on in the center port, instructing the rats to nose poke there to initiate a trial. They were required to keep their noses in the center port until the center LED offset (nose fixation). Broken fixation trials were ignored in all analyses. During the first 1 s of nose fixation, a Pro or Anti sound was played (clearly distinguishable FM modulated sounds) to indicate the current task, followed by a 500-ms silent delay when rats had to remember the current task while maintaining nose fixation. The center LED was then turned off, allowing the animal to withdraw from the center port. The withdrawal would trigger either a left or right LED to turn on as the target stimulus, which remained on until rats poked into one of the side ports. Response time (RT) is defined as the time from target onset until side poke. On a Pro trial, rats were rewarded for orienting towards the side LED; on an Anti trial, rats were rewarded for orienting away from the side LED and into the port without light. A correct choice was rewarded by 24 µl of water; and an incorrect choice resulted in a loud sound, no reward, and a short time-out. To ensure that all sub-trial optogenetic inactivation conditions have the same duration for laser stimulation (750 ms), all rats implanted with optical fibers were trained on a modified version of the behavior where the task cue period and the delay period both lasted 750 ms instead of the 1-s cue period and the 500-ms delay period as in the original design.

In all recording and inactivation sessions, rats performed alternating blocks of Pro and Anti trials, where block switches occurred within single sessions, after a minimum of 15 trials per block, and when a local estimate of performance (over the last ten trials in this block) reached a threshold of 70% correct. Detailed training procedures and codes can be found in a previous report[10].

Recordings. Rats were implanted with custom-made movable microdrives and recordings were made with platinum-iridium tetrodes[46]. To target the prelimbic (PL) area of PFC (+3.2 anteroposterior [AP] mm, ±0.75 mediolateral [ML] mm from bregma), tetrodes were initially positioned at ~1.5 mm below brain surface and were advanced daily during recording sessions to sample different neurons. To target the intermediate and deep layers of the SC (-6.8 AP mm, ±1.8 ML mm), tetrodes were initially positioned at ~3 mm below brain surface and advanced daily. Electrode placements were confirmed with histology. Four rats had both PL and SC implants (same hemisphere), 2 rats had a PL implant only, and 1 rat had an SC implant only. The choice of recording area and hemisphere side was assigned randomly for each rat.

Analysis of neural data. Spike sorting was done manually using SpikeSort3D (Neuralynx), and only isolated single units were included in the following analyses. In order to perform analyses on the neural population, we only analyzed neurons recorded for a sufficient number of trials. More specifically, we only analyzed neurons for which we had collected responses during at least 25 correct trials for each of the four possible task conditions (Pro-Right, Pro-Left, Anti-Right, Anti-Left). This resulted in the analysis of 193 neurons (out of 215) in SC, and 291 neurons (out of 331) in PFC. The response of each neuron was quantified by counting the number of spikes in 250ms-wide bins. In all analyses, the response was aligned to the time when the target stimulus appeared (i.e. the time of withdrawal from the center port). The temporal gap between the fixation offset and target stimulus onset was controlled by animals and thus variable on each trial. On average, rats withdrew from the center port 127 ms after fixation offset. Therefore, in all figures, we indicate the start of the delay period (end of task cue presentation) 0.627 s before target stimulus onset (500 ms delay + 127 ms), and the start of task cue presentation

at 1.627 s before target onset (1 s of task cue presentation before the delay). Unless otherwise noted, in all figures the time scale is causal, i.e. the value at time 0 refers to the neural activity in a time bin between -250 ms and 0 ms.

Quantification of single neuron selectivity. The amount of information encoded by a single neuron about a task variable was measured at each time point using d', defined as the difference in the number of spikes fired in response to two generic task conditions (here named as A and B), normalized by the square root of the pooled variance: $d' = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}}$, where μ_A indicates the mean spike count in response to condition A, μ_B indicates the mean spike count in response to condition B, ${\sigma^2}_A$ indicates the variance across trials of the spike count in response to condition A, and ${\sigma^2}_B$ indicates the variance across trials of the spike count in response to condition B.

Information about the task rule (Pro/Anti d') was computed by comparing the responses during Pro trials and the responses during Anti trials (with positive d' indicating Pro-preference). Information about the rat's choice (Choice d') was computed by comparing the responses during trials that resulted in an orienting movement contralateral to the recorded neuron, and trials that resulted in an ipsilateral orienting movement (with positive d' indicating Contra-preference; Extended Data Figure 3).

The threshold above which a d' value was considered significantly different than 0 was computed based on the pairwise t-test between the two conditions, using a p-value of 0.05. In Figure 2a, d' significance at each time point was computed using a shuffling procedure to correct for multiple comparisons, where the d' at each time point was recomputed 100 times after randomly shuffling the labels of Pro and Anti trials, and the 95th percentile of the resulting overall distribution of shuffled d's was used as the significance threshold.

Single neuron selectivity about the task rule was used to define two distinct classes of neurons (Fig. 2a). "Cue neurons" were defined as those with peak Pro/Anti d' at a time while the task cue was still being presented. "Delay/Choice neurons" were defined as those with peak Pro/Anti d' at times after the task cue was no longer present. Neurons whose Pro/Anti d' was never significantly higher than 0 were excluded from both groups.

Within the class of "Delay/choice neurons", we used single neuron selectivity about the choice in the first time bin after stimulus presentation (i.e. from 0 to 250 ms) to further subdivide these cells into two groups (Fig. 3). "Contra neurons" had a significantly higher response to contralateral stimulus, whereas "Ipsi neurons" had a significantly higher response to an ipsilateral stimulus.

Population-level decoding analysis. To determine the amount of task-relevant information available in the SC and PFC neural populations at each time point, we performed a series of cross-validated linear classification analyses[20]. For each analysis, we considered the spike count responses of a population of N neurons to a task condition as a population "response vector" **x**, and we randomly assigned 60% of the recorded trials (30 trials) as the training set, and the remaining 40% of the trials (20 trials) as the test set. The training set was used to compute the linear hyperplane that would optimally separate the population response vectors corresponding to two different task conditions (e.g. Pro trials vs Anti trials). This linear readout can also be written as $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ where **w** is the N-dimensional vector of weights applied to each of the neurons, and b is a scalar threshold. The classification of a test response vector **x** was then assigned depending on the sign of f(**x**) , and the performance was computed as the fraction of correct classifications over 500 resampling iterations. Because some of the neurons were recorded in different sessions, trials were always shuffled on each iteration to destroy any artificial trial-by-trial correlations. The hyperplane and threshold were computed using a Support Vector Machine algorithm using the LIBSVM library (https://www.csie.ntu.edu.tw/~cjlin/libsvm).

When comparing the classification performances for neural populations with different numbers of neurons, we randomly resampled identical numbers of neurons without replacement on each iteration. Because the overall

average firing rate was higher in SC than in PFC, we tested whether matching firing rates was sufficient to explain the classification result (Extended Figure 4a), by removing single spikes at random from the SC dataset until the average firing rates were matched, and by performing again the classification analysis on the equalized SC population. To produce an estimate of the number of PFC neurons necessary to match performances in SC (Extended Figure 4b), we adopted an analytical approach to estimate classification performances based on the distribution of d' in a neural population[47]. More specifically, we quantified the Normalized Euclidean Distance (NED) between two conditions in a neural population as the square root of the sum of the squared d's across all neurons: $NED = \sqrt{\sum d'^2}$. This quantity can be used to estimate linear classification performances, under the Gaussian assumption, as[48]: $Perf = 1 - H\left(\frac{k_{eff} \cdot NED}{2}\right)$, where H is the complementary error function, and k_{eff} is an efficiency factor that accounts for the inability of the classifier to extract all the available information (e.g. due to limited training data). Before applying this formula, d's had to be corrected for their intrinsic positive bias[49]. Because the NED grows with the square root of the total number of neurons in a population, we could then estimate the classification performance for a neural population of arbitrary size M as: $Perf_N = 1 - H\left(\frac{k_{eff} \cdot NED \cdot \sqrt{M}}{2 \cdot \sqrt{N}}\right)$, where N indicates the actual size of the population for which NED was computed. Using the same approach, we also tested how measurements of latency in the rise of classification performance (see below) depended on the total number of neurons in a neural population (Extended Figure 4c).

When classification analyses were used to compare performances during correct and error trials (Fig. 2b), we always trained the classifier using correct trials, and we tested the classifier using either correct or error trials. The number of trials used for testing was limited by the neuron with the fewest number of error trials per condition (9 trials).

To compute the latency of the rise in choice classification performance for different neural populations (Fig. 2c), we evaluated the average time after the appearance of the target stimulus necessary for the population readout to reach a fixed threshold (correct performance >65%)[47]. More specifically, on each iteration of the resampling procedure we computed the classification performances for each time point, we smoothed the resulting curve by averaging the value for 5 neighboring time points, and we noted the time point where the curve crossed the performance threshold. We computed the mean and the standard error of the latency as the mean and standard deviation of these values.

To compute the significance of differences in the magnitude (or latency) of population performances, we adopted a bootstrap approach based on our resampling procedure[50]. More specifically, we first evaluated the average performance (or latency) across all iterations for the two populations, and we then computed the p-value as the fraction of iterations in which, by chance, the value for the population with the lower average was above the value for the population with the higher average.

Optical fiber construction, virus injection and fiber implantation. Chemically sharpened optical fibers (50/125 um LC-LC duplex fiber cable, http://www.fibercables.com) were prepared as previously described[22]. To ensure the distance between the two optical fibers was the distance between bilateral SC (3.6 mm), we inserted two metal cannulae into a plastic template and guided the optical fibers through the cannulae, which were 3.6 mm apart (Fig. 4a).

Basic virus injection techniques were identical to those described previously[22]. At the targeted coordinates (SC, -6.8 AP mm, ±1.8 ML mm from bregma), two injections of 9.2 nl AAV virus (AAV5-CaMKIIα-eYFP-eNpHR3.0 for inactivations, 9 rats; AAV5-CaMKIIα-eYFP for controls, 3 rats) were made every 100 um in depth starting 3.5 mm below brain surface for 1.5 mm. Four additional injection tracts were completed, one 500 um anterior, posterior, medial, and lateral from the central tract. A total of 1.5 µl of virus was injected over the course

of 30 minutes. Chemically sharpened bilateral SC fiber implant was lowered down the central injection track, with the tip of each fiber positioned at 4.4 mm below brain surface to target the center of SC's intermediate and deep layers. Training was resumed 5 days post-surgery. Virus expression was allowed to develop for 8 weeks before behavioral testing began.

Optogenetic inactivation and analysis. For each inactivation session, animals' implants were connected to a 1m patch cable connected to a fiber rotary joint (Princetel) mounted above the behavioral chamber. A 200 mW 532 nm laser (OEM Laser Systems) was then connected to deliver constant light at 25 mW per site, with a < 5 mW difference between the left and right SC. Laser illumination occurred on 25% randomly chosen trials in each behavioral session. Different optogenetic conditions (3-s full-trial inactivation, 750-ms task cue, 750-ms delay, or 750-ms choice period inactivation) were randomly interleaved for all sessions to control for behavioral fluctuations across days.

Behavioral changes due to optogenetic inactivation were quantified as the performance difference between inactivation (laser) trials and control (no-laser) trials from the same sessions. These results are then compared to YFP control data. For each session, we calculated the baseline error rate or RT for Pro and Anti control trials and subtracted that mean value from the performance on individual inactivation trials. After obtaining the normalized changes in performance due to inactivation for individual sessions, we concatenated trials across all sessions and all rats, and computed the mean and s.e.m. across trials. Nonparametric bootstrap procedures or permutation tests were used to compute significance values (shuffled 5000 times). All rats were included in the full-trial inactivation analyses. For sub-trial inactivation analyses, we only included the rats (8/9) that had significant full-trial effects.

Acute characterization of optogenetic effects. To measure the effects of optogenetic inactivation on neural activity, acute recordings of infected SC neurons were performed in anesthetized rats (Fig. 4b). An etched fiber optic and sharp tungsten electrode (0.5 or 1.0 MΩ) were independently advanced to the center of the infected area. For each neuron tested, baseline neural activity was recorded for 2 s, followed by 8 s of laser stimulation at 25 mW, and another 2 s of post-stimulation recording, repeated for >10 times. We observed that the onset and offset of optogenetic inactivation of neural activity was within 50 ms of laser onset and offset (Fig. 4b).

Model setup. Our model consists of four dynamical units, each unit had an external(V_i) and internal (U_i) variable. The relationship between the internal and external variables is given by:

$$V_i(t) = (0.5 \cdot tanh((U_i(t) - \theta)/\beta) + 0.5) \cdot \eta(t)$$

Here $\eta(t)$ is the optogenetic inactivation fraction, which tells us the fraction of this unit's output that is silenced by optogenetic inactivation in a time-dependent fashion (1 = no optogenetic inactivation). $\beta = 0.5$ controls the slope of the input-output relationship, and $\theta = 0.05$ controls the midpoint of the input-output function. The internal variables had dynamical equations:

$$\tau \cdot dU_i/dt = -U_i + W \cdot V_i + input + \sigma \cdot dW$$

Where W is the network weight matrix, *input* is the external input into the network, $\tau = 0.09\ s$ is a fixed time constant for each unit, and $\sigma \cdot dW$ is gaussian noise with amplitude given by the parameter σ. W was parameterized by four parameters that controlled the self-weights sW, the horizontal weights hW between the two Pro units and between the two Anti units, the vertical weights vW between the two right units and between the two left units, and the diagonal weights dW between Pro-R/Anti-L and between Pro-L/Anti-R. The external input into the network was given by:

$$input = E_{constant} + E_{Pro-bias} + E_{rule} + E_{choice-period} + E_{light}$$

$E_{constant}$ is constant excitation to all units. Parameter $E_{Pro-bias}$ is constant excitation to both Pro units, but not to the Anti units. E_{rule} is the rule input, which is only active during the rule and delay periods, and not during the choice period. On Anti trials, the two Anti units get rule input $E_{Anti-rule}$ and on Pro trials, the two Pro units get

rule input $E_{Pro-rule}$. Parameter $E_{choice-period}$ is excitation to all units only during the target period when a light cue is presented and animals are free to choose. Parameter $E_{choice-period}$ is excitation to both units on the side (L vs R) activated by the light cue, when the cue is active. Each trial was simulated numerically used the forward-euler method with time step dt=0.024s, which we found to balance accuracy and computational speed. The duration of each trial was rule + delay period = 1.2s, target period = 0.3s, post_target_period =0.3s. Individual trials of the same trial type are differentiated by the noise samples generated by the additive gaussian noise process.

Model cost function.

The cost function has two terms $C = C_1 + C_2$, C_1 penalizes model performance that deviates from the target performance, C_2 penalizes weak model choices where the output units are close together. Below we will describe C_1 and C_2 respectively.

C_1 *Term:* To read out the models choice on a given trial, we could ask if $V_{Pro-R} > V_{Pro-L}$. However, this creates a discontinuity in the cost function if a small change in a parameter causes the decision to flip. In order to use powerful optimization tools like automatic differentiation, we wanted the cost function to be fully differentiable. Therefore, each model choice was recast as the probability of a choice by passing the unit outputs through a tanh() function with a sensitivity given by a fixed parameter θ_1. For a Pro trial, the probability of a correct choice was given by: HitP =$0.5 * (1 + tanh((V_{Pro-R} - V_{Pro-L})/\theta_1))$, and for an anti trial: HitA = $0.5 * (1 + tanh((V_{Pro-L} - V_{Pro-R})/\theta_1))$. For each trial type, i, we defined a target hit percentage, and penalized the difference between the target hit percentage and the average hit percentage from the model across all trials. The overall cost from this first term was the sum across trial types. $C_1 = \sum_i \left(\overline{hitP_i} - TargetHitP_i\right)^2$.

C_2 *Term:* The tanh() makes the cost function differentiable, but encourages the model to reach the target hit percentage on *every trial*, rather than making strong choices on each trial, some right and some wrong, that average to the target hit percentage. To prevent this degenerate solution, we introduced a second cost term that penalizes weak choices where the activation of the two Pro units are close. For a Pro trial: $C_2 = -\beta_c \left(tanh\left((V_{Pro-R} - V_{Pro-L})/\theta_2\right)\right)^2$, and for an Anti trial: $C_2 = -\beta_c \left(tanh\left((V_{Pro-L} - V_{Pro-R})/\theta_2\right)\right)^2$. θ_2 is a fixed parameter that controls the sensitivity of this term, and β_c is a fixed parameter controlling the strength of this term. We used the fixed parameter values $\theta_1 = 0.05$, $\theta_2 = 0.15$, $\beta_c = 0.001$.

Model Optimization.

We initialized many different model solutions with random parameter values, and a random seed for the random number generator to generate unique noise for each model solution. For each initialization we minimized the cost function using constrained parabolic minimization.

Constrained Parabolic Minimization: The minimization starts by creating a local search radius, which restricts the scope of the search on each step. At each step, the algorithm approximates the cost function locally using the hessian matrix, and gradient vector, which defines a 12 dimensional parabolic surface. The minimization takes a step in the direction that minimizes the cost on this parabola subject to the constraint that the step length equals the search radius. If the resulting step would increase the cost function the step is not taken, the search radius is reduced, and another step is attempted. As the search radius becomes smaller, this method converges to gradient descent.

Two Stage Optimization: For each parameter initialization, an initial minimization was done using 50 trials/condition. If this initial minimization passed a set of criteria then a further minimization was done using 1600 trials/condition. The initial criteria were used to prevent long optimizations on model solutions that were

performing very badly. The initial criteria were that performance on Pro trials was greater than Anti trials, and Anti performance on delay-period opto trials was worse than control or choice-period opto trials. The final minimization terminated after 1000 iterations, or when a step in parameter space reduced the cost function by less than 1e-12. When the minimization ended, if the final cost was below a threshold of -0.0002 we accepted the final parameter values as a model solution. We ended up with $N = 354$ unique model solutions.

Model analysis. To examine the space of model solutions, we clustered each model solution based on the dynamics of their simulated units. We simulated 200 trials for each model solution for each trial type (total 6 = Pro/Anti x control/delay-period opto/choice-period opto). Then, we computed the average trajectory for each unit in the model on correct and incorrect trials for each trial type. The average trajectories were concatenated into a model response vector (length M = 4 units x hit/miss x Pro/Anti x control/delay/choice opto x T timesteps). We created response matrix (R) of response vectors for all model solutions. (Size N x M). We used the Singular Value Decomposition to factor $R = U * S * V'$. The orthonormal matrices U and V' are the directions of greatest variance in R across model solutions (U), and across time points (V'). The column space of U gives us the weights for each model solution onto the set of temporal basis vectors in the row space of V'. We clustered model solutions by looking at the leading columns of U, which explain the most variance across model solutions. To determine the correct number of clusters, we considered the reduced subspace formed by the 3 leading columns of U (accounting for 68% of the variance), and we measured the Bayes Information Criterion (BIC) associated with fitting different numbers of Gaussian Mixtures[51]. The number of clusters with lowest BIC was associated with 7 clusters, and this number was used to perform a k-means clustering, using k-means++ for the initialization of the cluster centers[52].

ACKNOWLEDGEMENTS

We thank K. Osorio and J. Teran for animal and laboratory support. This work was funded by Howard Hughes Medical Institute. C.A.D. was supported by a Howard Hughes Medical Institute predoctoral fellowship. C.A.D. and M.P. are currently supported by the Simons Collaboration on the Global Brain postdoctoral fellowship.

AUTHOR CONTRIBUTIONS

C.A.D. collected electrophysiological and optogenetics data. M. P. and C.A.D. analyzed electrophysiological data. C.A.D. analyzed the optogenetics data. M.P., A.T.P., C.D.B. and A.J.R. generated and analyzed modeling results. A.A. and C.D.K. carried out the acute optogenetics experiments. C.A.D., J.C.E. and C.D.B. conceived the project. C.A.D., M.P., A.T.P. and C.D.B. wrote the paper. J.C.E. and C.D.K. played an advisory role on electrophysiological and optogenetics experiments respectively. C.D.B. was involved in all aspects of experimental design and data analysis.

REFERENCES

1. Real, E., Asari, H., Gollisch, T. & Meister, M. Neural Circuit Inference from Function to Structure. *Curr. Biol.* **27,** 189–198 (2017).
2. Goldman, M. S., Golowasch, J., Marder, E. & Abbott, L. F. Global structure, robustness, and modulation of neuronal models. *J. Neurosci.* **21,** 5229–5238 (2001).
3. Prinz, A. A., Bucher, D. & Marder, E. Similar network activity from disparate circuit parameters. *Nat. Neurosci.* **7,** 1345–1352 (2004).
4. Gutierrez, G. J., O'Leary, T. & Marder, E. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron* **77,** 845–858 (2013).
5. Fisher, D., Olasagasti, I., Tank, D. W., Aksay, E. R. F. & Goldman, M. S. A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron* **79,** 987–1000 (2013).
6. Transtrum, M. K. *et al.* Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* **143,** 010901 (2015).
7. Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. Automatic differentiation in machine learning: a survey. *arXiv [cs.SC]* (2015).
8. Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307,** 1121–1124 (2005).
9. Wong, K.-F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26,** 1314–1328 (2006).
10. Duan, C. A., Erlich, J. C. & Brody, C. D. Requirement of Prefrontal and Midbrain Regions for Rapid Executive Control of Behavior in the Rat. *Neuron* **86,** 1491–1503 (2015).
11. Munoz, D. P. & Everling, S. Look away: the anti-saccade task and the voluntary control of eye movement. *Nat. Rev. Neurosci.* **5,** 218–228 (2004).
12. Everling, S. & DeSouza, J. F. X. Rule-dependent activity for prosaccades and antisaccades in the primate prefrontal cortex. *J. Cogn. Neurosci.* **17,** 1483–1496 (2005).
13. Weiler, J. & Heath, M. Task-switching in oculomotor control: unidirectional switch-cost when alternating between pro- and antisaccades. *Neurosci. Lett.* **530,** 150–154 (2012).

14. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497,** 585–590 (2013).

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

16. Johnston, K., DeSouza, J. F. X. & Everling, S. Monkey prefrontal cortical pyramidal and putative interneurons exhibit differential patterns of activity between prosaccade and antisaccade tasks. *J. Neurosci.* **29,** 5516–5524 (2009).

17. Everling, S., Dorris, M. C., Klein, R. M. & Munoz, D. P. Role of primate superior colliculus in preparation and execution of anti-saccades and pro-saccades. *J. Neurosci.* **19,** 2740–2754 (1999).

18. Felsen, G. & Mainen, Z. F. Midbrain contributions to sensorimotor decision making. *J. Neurophysiol.* **108,** 135–147 (2012).

19. Chan, J. L., Koval, M. J., Johnston, K. & Everling, S. Neural correlates for task switching in the macaque superior colliculus. *J. Neurophysiol.* **118,** 2156–2170 (2017).

20. Pagan, M., Urban, L. S., Wohl, M. P. & Rust, N. C. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat. Neurosci.* **16,** 1132–1139 (2013).

21. Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. & Brody, C. D. Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements. *Neuron* **88,** 367–377 (2015).

22. Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520,** 220–223 (2015).

23. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532,** 459–464 (2016).

24. Robinson, D. A. Eye movements evoked by collicular stimulation in the alert monkey. *Vision Res.* **12,** 1795–1808 (1972).

25. Guitton, D., Crommelinck, M. & Roucoux, A. Stimulation of the superior colliculus in the alert cat. *Exp. Brain Res.* **39,** 63–73 (1980).

26. Dean, P., Mitchell, I. J. & Redgrave, P. Contralateral head movements produced by microinjection of glutamate into superior colliculus of rats: evidence for mediation by multiple output pathways. *Neuroscience*

24, 491–500 (1988).

27. Goffart, L., Hafed, Z. M. & Krauzlis, R. J. Visual fixation as equilibrium: evidence from superior colliculus inactivation. *J. Neurosci.* **32,** 10627–10636 (2012).
28. May, P. J. The mammalian superior colliculus: laminar structure and connections. *Prog. Brain Res.* **151,** 321–378 (2006).
29. Wolf, A. B. *et al.* An integrative role for the superior colliculus in selecting targets for movements. *J. Neurophysiol.* **114,** 2118–2131 (2015).
30. Guitton, D., Buchtel, H. A. & Douglas, R. M. Frontal lobe lesions in man cause difficulties in suppressing reflexive glances and in generating goal-directed saccades. *Exp. Brain Res.* **58,** 455–472 (1985).
31. Pierrot-Deseilligny, C., Rivaud, S., Gaymard, B. & Agid, Y. Cortical control of reflexive visually-guided saccades. *Brain* **114 (Pt 3),** 1473–1485 (1991).
32. Fukushima, J. *et al.* Disturbances of voluntary control of saccadic eye movements in schizophrenic patients. *Biol. Psychiatry* **23,** 670–677 (1988).
33. Hutton, S. B. & Ettinger, U. The antisaccade task as a research tool in psychopathology: a critical review. *Psychophysiology* **43,** 302–313 (2006).
34. Pierrot-Deseilligny, C. *et al.* Decisional role of the dorsolateral prefrontal cortex in ocular motor behaviour. *Brain* **126,** 1460–1473 (2003).
35. Lo, C.-C. & Wang, X.-J. Conflict Resolution as Near-Threshold Decision-Making: A Spiking Neural Circuit Model with Two-Stage Competition for Antisaccadic Task. *PLoS Comput. Biol.* **12,** e1005081 (2016).
36. Condy, C., Wattiez, N., Rivaud-Péchoux, S., Tremblay, L. & Gaymard, B. Antisaccade deficit after inactivation of the principal sulcus in monkeys. *Cereb. Cortex* **17,** 221–229 (2007).
37. Wegener, S. P., Johnston, K. & Everling, S. Microstimulation of monkey dorsolateral prefrontal cortex impairs antisaccade performance. *Exp. Brain Res.* **190,** 463–473 (2008).
38. Pierrot-Deseilligny, C., Rosa, A., Masmoudi, K., Rivaud, S. & Gaymard, B. Saccade deficits after a unilateral lesion affecting the superior colliculus. *J. Neurol. Neurosurg. Psychiatry* **54,** 1106–1109 (1991).
39. Everling, S. & Johnston, K. Control of the superior colliculus by the lateral prefrontal cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20130068 (2013).

40. Johnston, K., Koval, M. J., Lomber, S. G. & Everling, S. Macaque dorsolateral prefrontal cortex does not suppress saccade-related activity in the superior colliculus. *Cereb. Cortex* **24,** 1373–1388 (2014).

41. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24,** 167–202 (2001).

42. Horwitz, G. D., Batista, A. P. & Newsome, W. T. Representation of an abstract perceptual decision in macaque superior colliculus. *J. Neurophysiol.* **91,** 2281–2296 (2004).

43. Zénon, A. & Krauzlis, R. J. Attention deficits without cortical neuronal deficits. *Nature* **489,** 434–437 (2012).

44. Basso, M. A. & May, P. J. Circuits for Action and Cognition: A View from the Superior Colliculus. *Annu Rev Vis Sci* **3,** 197–226 (2017).

45. Krauzlis, R. J., Lovejoy, L. P. & Zénon, A. Superior colliculus and visual spatial attention. *Annu. Rev. Neurosci.* **36,** 165–182 (2013).

46. Erlich, J. C., Bialek, M. & Brody, C. D. A cortical substrate for memory-guided orienting in the rat. *Neuron* **72,** 330–343 (2011).

47. Pagan, M. & Rust, N. C. Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. *J. Neurosci.* **34,** 11067–11084 (2014).

48. Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95,** 3633–3644 (2006).

49. Pagan, M. & Rust, N. C. Quantifying the signals contained in heterogeneous neural responses and determining their relationships with task performance. *J. Neurophysiol.* **112,** 1584–1598 (2014).

50. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (1993).

51. Claeskens, G. & Hjort, N. L. The Bayesian information criterion. in *Model Selection and Model Averaging* 70–98

52. Arthur, D. & Vassilvitskii, S. k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM* (2007).

Figure 1

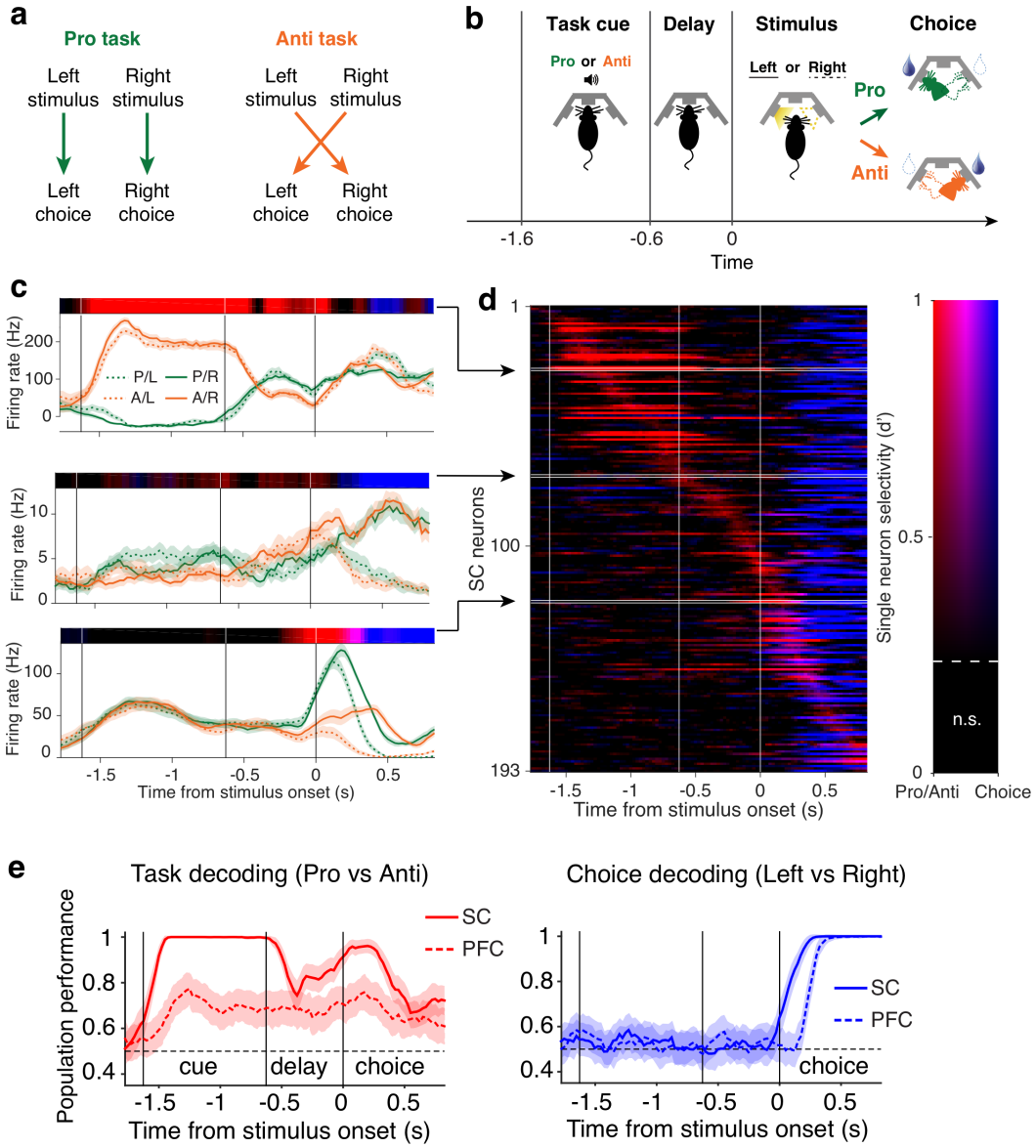


Figure 1 | SC and PFC populations contain task and choice information during rapid sensorimotor task switching. **a**, Rules for the Pro and Anti task contexts. In the Pro task, rats should orient *toward* a lateralized stimulus (left or right) for reward; in the Anti task, rats should orient *away* from the stimulus for reward. Trained rats can switch between these two known task contexts from one trial to the next. **b**, Rats nose poke in the center port to initiate each trial and keep fixation during the task cue (Pro or Anti sound) and delay periods. After the delay, the animal is allowed to withdraw from the center port, and a lateralized light (left or right) is turned on to indicate the stimulus location. Rats then poke into one of the side pokes for reward. **c**, Peri-stimulus time histogram (PSTH) for 3 example SC neurons on Pro-Go-Right (green solid), Pro-Go-Left (green dashed), Anti-Go-Right (orange solid), and Anti-Go-Left (orange dashed) trials. PSTHs are aligned to stimulus onset. Top, task (red) and choice (blue) selectivity as a function of time for each neuron. **d**, Information encoding matrix of the SC population. Each row of the matrix represents the d' of a single neuron as a function of time. The intensity of the color represents how "informative" a neuron is, and the RGB values are associated with different types of information (Pro/Anti, red; choice, blue; mixed, purple). Neurons are sorted by the timing of their peak Pro/Anti d'. d' that are not significant (n.s.) are set to 0. **e**, Evolution of classification performance over time in the SC (solid) and PFC (dashed) population. Left, mean and s.e.m. performance for linear classification of correct Pro versus Anti trials. Spikes are aligned to stimulus onset, and counted over windows of 250 ms with 25-ms shifts between neighboring windows. Note that performance is plotted over the right edge of the window (causal). Right, classification performance to linearly separate Go-Left versus Go-Right trials, similar to the left panel.

Figure 2

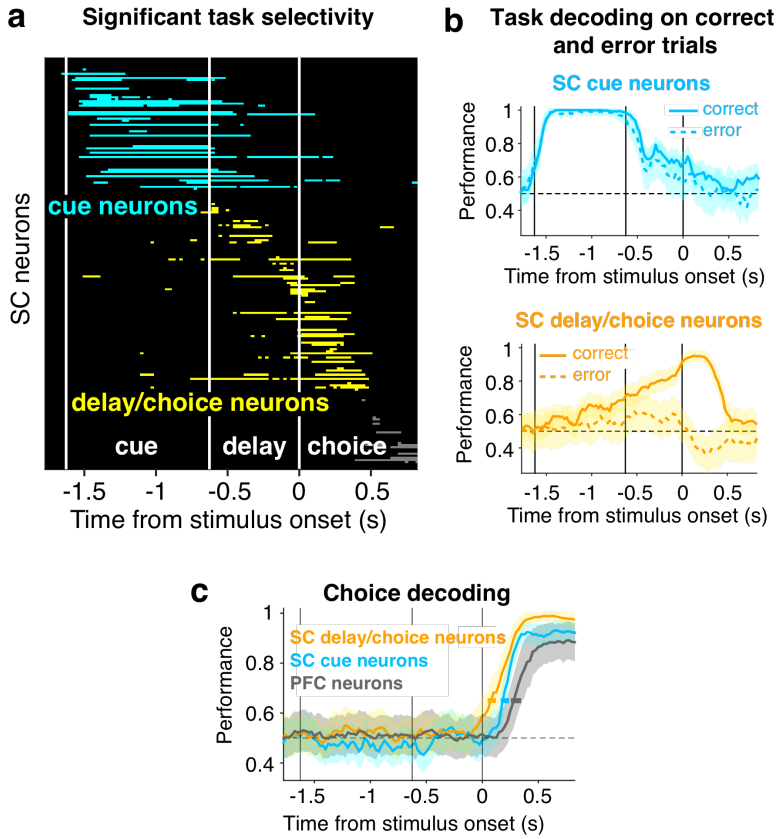


Figure 2 | Distinct roles of SC subpopulations. a, Timing of significant Pro/Anti selectivity (d') for all SC neurons, sorted by peak d'. Significance threshold was determined by shuffled data. We separated SC neurons into two groups based on the timing of their Pro/Anti selectivity. "Cue neurons" (cyan, n=29) differentiated between Pro and Anti trials most strongly during the auditory cue; "delay/choice neurons" maintained task selectivity most strongly when the auditory cue was no longer present (yellow, n=45). **b**, Performance of task decoding on correct versus error trials. Linear classifiers trained on correct trials were tested for separate correct trials (solid) or error trials (dashed). The representation of task context by delay/choice neurons was disrupted on error trials whereas such information in the cue neurons did not differentiate between correct and error trials. **c,** Choice decoding performance of SC subpopulations and PFC neurons (n=29 to match number of cue neurons, see Methods). Choice information emerged first in SC delay/choice neurons. Shaded areas (vertical error bars) indicate s.e.m. of decoding accuracy for each population across time. Horizontal error bars represent s.e.m. of the timing of reaching 0.65 decoding accuracy for each population.

Figure 3

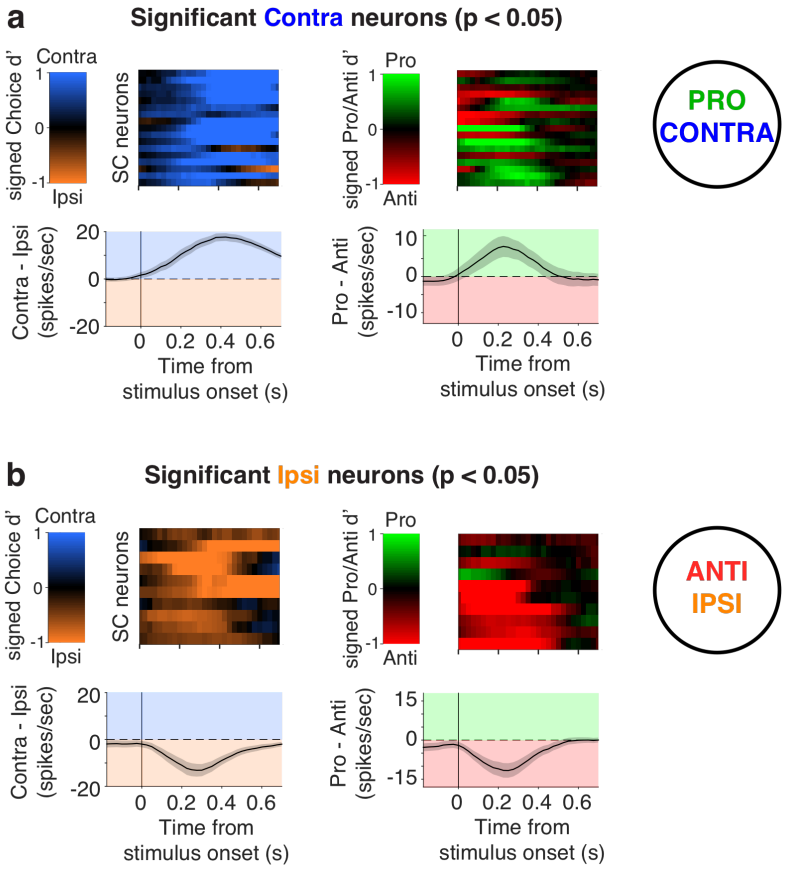


Figure 3 | A relationship between task and choice encoding around stimulus onset, suggesting two groups of neurons. a, Neurons selected as having a significantly greater firing rate on trials when the animal's choice is to orient Contralaterally (n = 17). Left top shows one neuron per row, with the color indicating the strength of firing rate difference between Contra- and Ipsi-orienting trials, as a function time relative to the visual stimulus onset. Left bottom shows firing rate difference (Contra-Ipsi) averaged over these neurons. Right panels are the same neurons as in the left panels, with each row corresponding to the same neuron as the one on the left, but now analyzed for Pro - Anti selectivity and firing rate. Contra-preferring neurons tend to be Pro-preferring neurons. **b,** as in panel a, but showing significantly Ipsi-preferring neurons (n = 10). These tend to also be Anti-preferring.

Figure 4

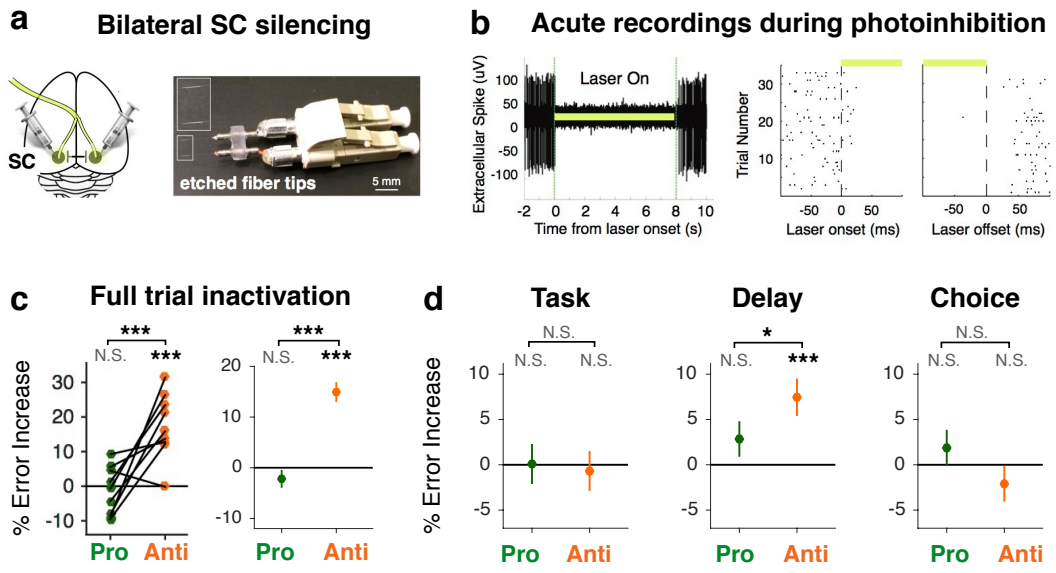


Figure 4 | SC delay activity is required for the Anti task. a, Experimental design for bilateral SC optogenetic inactivation. Left, a schematic that indicates virus infection and laser stimulation in the SC on both hemispheres. Right, an example of the optical fiber implant. The taper of each fiber is chemically sharpened to be approximately 2 mm long for stronger and more unified light delivery. The distance between the two fibers are constructed to be exactly 3.6 mm to target bilateral SC. **b**, Physiological confirmation of optogenetic inactivation effect in an anesthetized animal. Left: acute extracellular recording of spontaneous activity in the SC expressing eNpHR3.0. Laser illumination period (8 s) is marked by the light green bar. Right: spike activity aligned to laser onset and laser offset over multiple trials. Note that the onset and offset of the inhibitory effect are on the scale of tens of milliseconds. **c**, Effect of full-trial inactivation of bilateral SC. Mean Pro (green) and Anti (orange) error rate increase due to SC inactivation for all individual rats (n=9, left) and across all trials (Pro=662 trials, Anti=615 trials, right). Left, each data point represents the mean effect across sessions for a single rat. Right, means and s.e.m. across trials (concatenated across all 60 sessions). **d**, Effect of sub-trial inactivations of bilateral SC on Pro and Anti error rate (mean and s.e.m. across trials from 102 sessions). Statistical comparison between Pro and Anti effects were computed using a permutation test, shuffled 5000 times. N.S. $p>0.05$; *$p < 0.05$; ***$p < 0.001$. Note that all types of inactivations were randomly interleaved for each session.

Figure 5

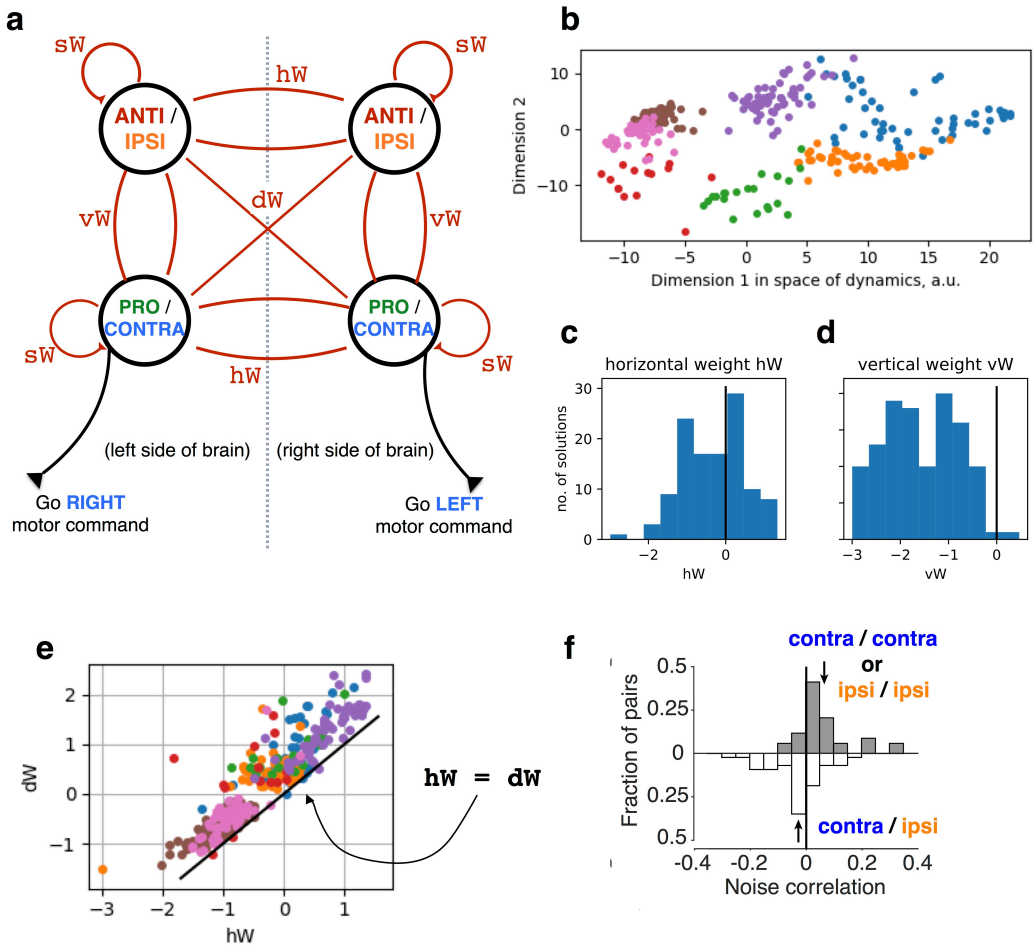
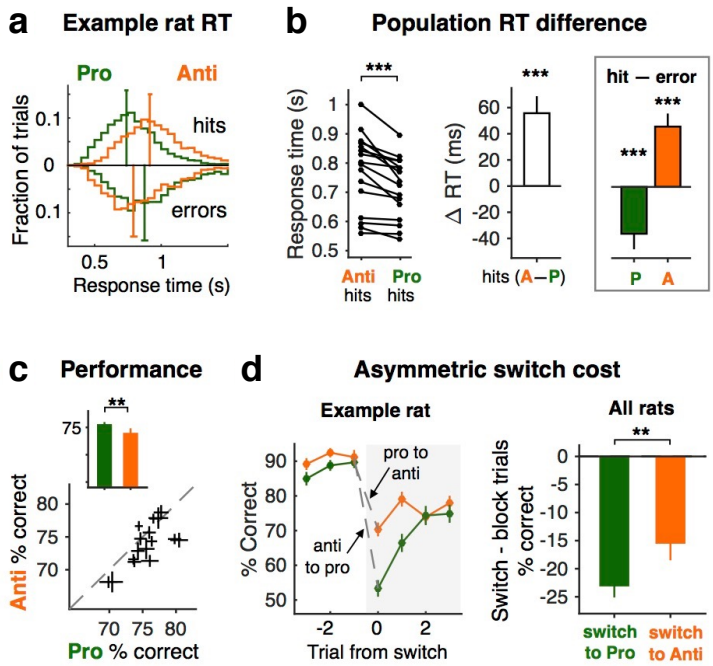


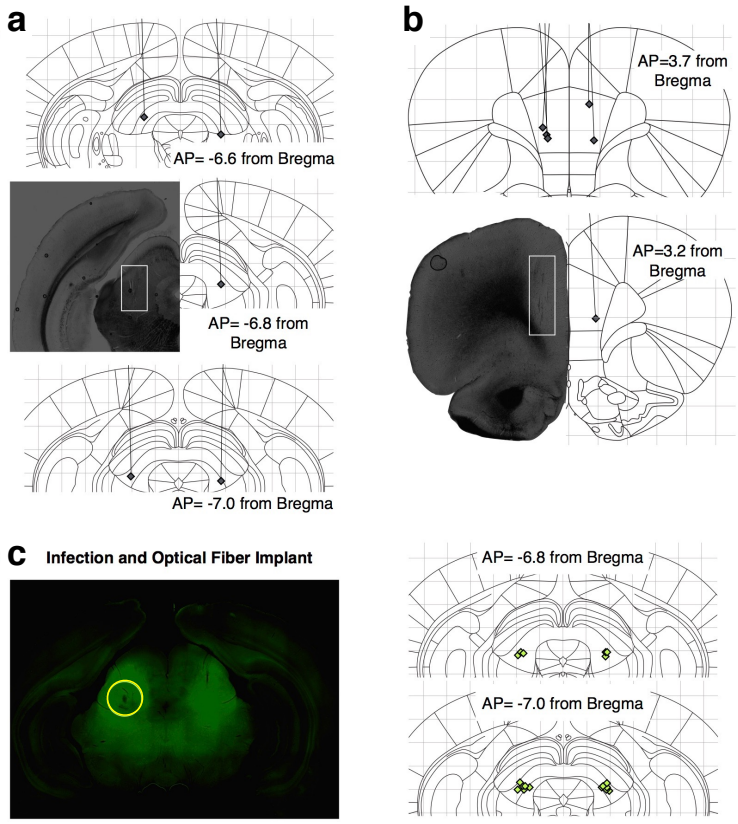
Figure 5 | Model and experimental data consistent with its predictions. a. Schematic of SC model, showing four units, each representing a population of SC delay/choice neurons. All connections are bidirectional. **b.** Projection of model solutions onto the 2D space that maximally explains variance in model dynamics across the set of model solutions. Individual dots are unique model solutions, color coding reflects the output of clustering on the model solutions. **c.** Histogram of horizontal weights in model solutions, showing a wide variety of values of mixed sign. **d.** Histogram of vertical weights in model solutions, showing a wide variety, but almost entirely negative. **e.** Scatter plot of horizontal weights against diagonal weight. Equality line added for reference. **f.** Trial-by-trial noise correlation between pairs of simultaneously recorded neurons on one side of the SC, calculated for within-group pairs of neurons (Contra/Contra or Ipsi/Ipsi, upper histogram) and between-group pairs (Contra/Ipsi). Noise correlation distribution for within-group pairs was significantly shifted above 0 (mean = 0.082 ± 0.02; $p<0.01$), whereas the between-group distribution was significantly shifted below 0 (mean = -0.052 ± 0.02; $p<0.05$), consistent with negative vW as predicted by the model. Arrows indicate the mean values.

Extended Data Figure 1



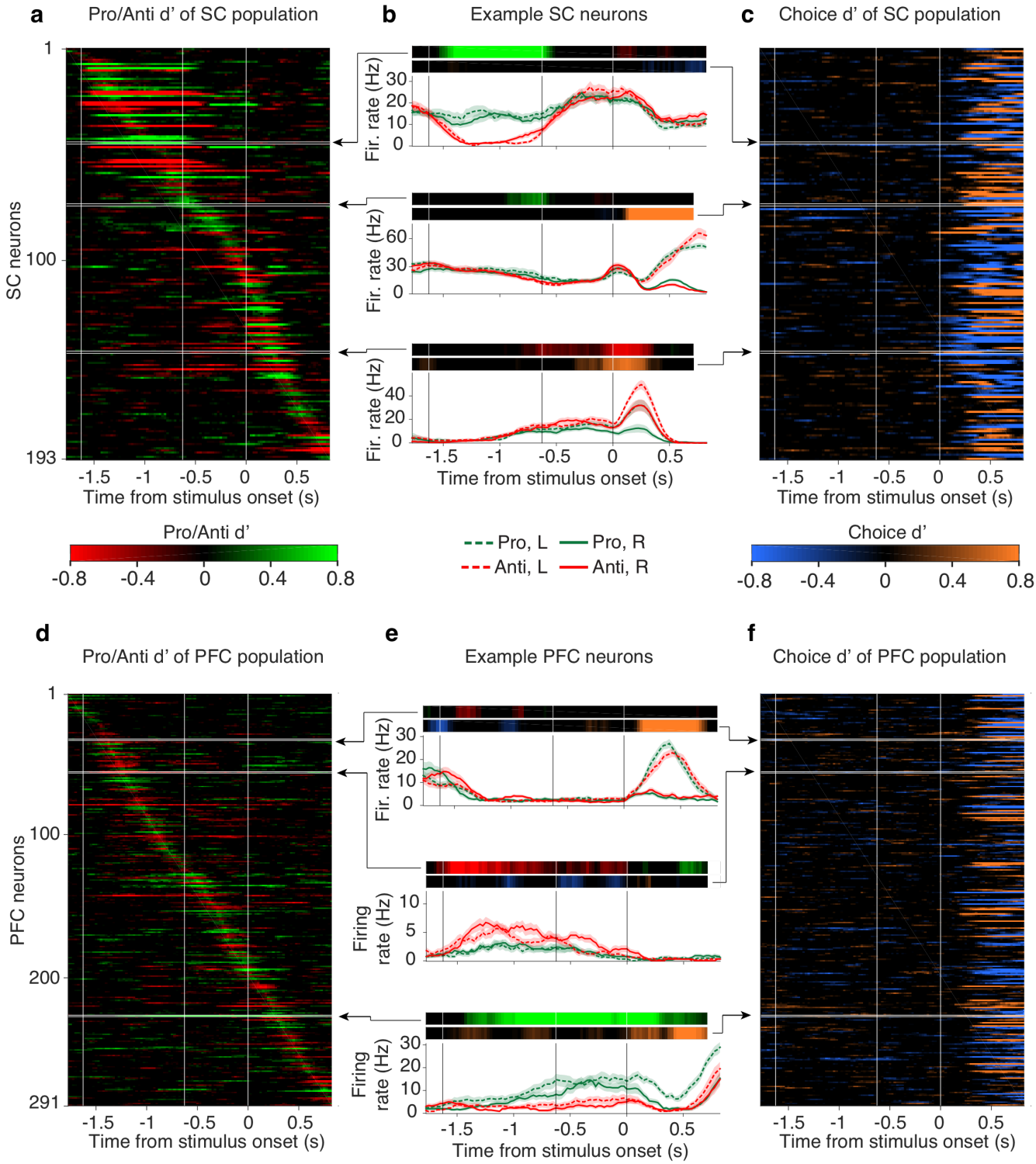
Extended Data Figure 1 | Post-surgery performance for implanted rats. Asymmetries between Pro and Anti response time (RT), accuracy, and task switch cost observed in implanted rats replicate those found in freely moving rats in . **a**, Normalized RT distributions of an example rat. Histograms of Pro and Anti RTs are shown here for hits (top) and errors (bottom). Each curve is normalized to have a total area of 1. Median RTs for Pro and Anti hits and errors are indicated by vertical bars; 95% confidence intervals across trials for each trial type are indicated by horizontal bars. **b**, RT summary of 16 individual rats (7 for neural recordings and 9 for optogenetic inactivation experiments). Left: median RTs for Anti hits and Pro hits for all rats. Right: RT difference between Pro and Anti, hits and errors, averaged across all rats. For each rat, the difference between median RTs of paired conditions was calculated. White bar shows the mean and SEM across rats for Anti hit RTs minus Pro hit RTs. Green bar shows Pro hit RTs minus Pro error RTs. Orange bar shows Anti hit RTs minus Anti error RTs. **c**, Pro and Anti performance for individual rats. Means and SEMs of Pro and Anti performance are computed over sessions for each rat and plotted against each other. Average Pro (green) and Anti (orange) performance across rats was plotted in the upper left corner. **d**, Switch cost asymmetry. Left: percent correct as a function of trial number relative to a task block switch for one example rat. Each data point is the mean and SEM across trials for Pro and Anti accuracy on three trials before and after the switch. Right: average accuracy switch cost for Pro trials and Anti trials across rats. **$p < 0.01$; ***$p < 0.001$.

Extended Data Figure 2



Extended Data Figure 2 | Histology for tetrode and optical fiber implantation. a, Histology for 5 rats with left or right SC tetrode implants. Gray diamonds indicate the final location of the tetrode tips. Lines indicate the tetrode tracks. **b**, Histology for 6 rats with left or right mPFC tetrode implants, similar to **a**. Seven rats were implanted with tetrode drives all together: 4/7 rats with both SC and mPFC drives; 2/7 rat with only mPFC drives; 1/7 rat with only SC drive. **c**, Histology for 9 rats with bilateral SC AAV virus infection and optical fiber implants. Left: example of AAV5-CaMKIIα-eYFP-eNpHR3.0 infection. Green fluorescence indicates the infection coverage. Yellow circle indicates the estimated spread of light stimulation based on previous acute recording experiments (Hanks et al., 2015). Right: green diamonds indicate the tips of etched optical fibers for all animals.

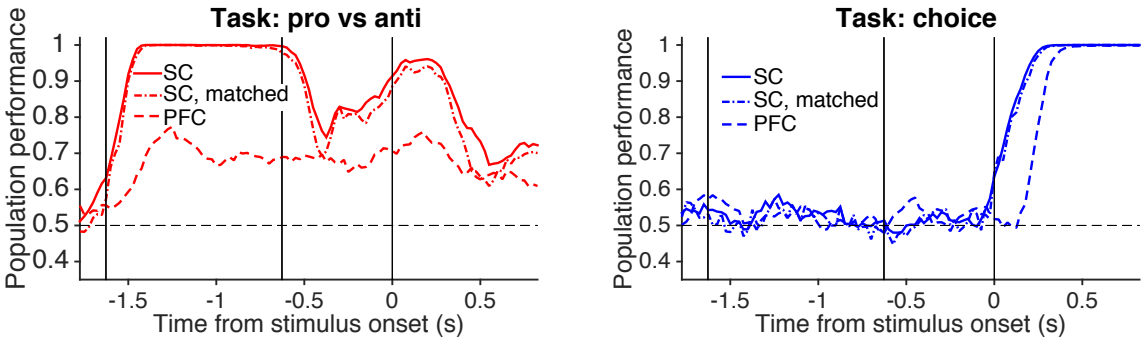
Extended Data Figure 3



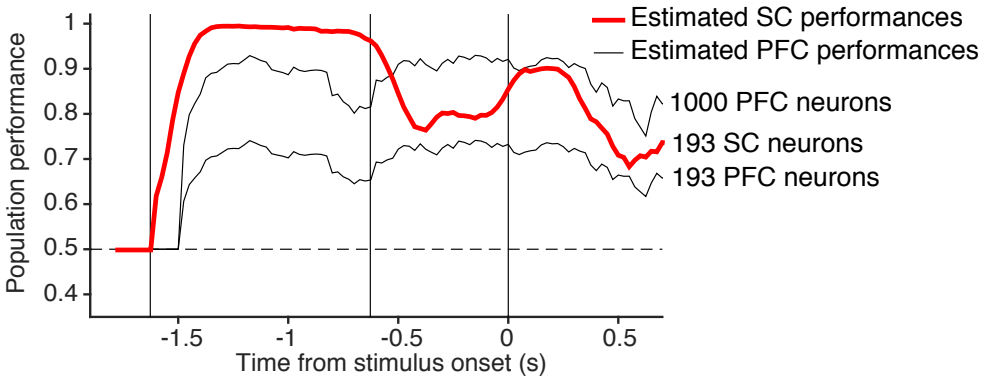
Extended Data Figure 3 | Heterogeneity of neural responses in SC and PFC. a, Matrix of Pro/Anti selectivity for the SC population. Each row of the matrix represents the Pro/Anti signed d' of a single neuron as a function of time. Neurons are sorted by the timing of their peak Pro/Anti absolute |d'|. **b**, Peri-stimulus time histograms (PSTHs) for 3 example SC neurons (same conventions as in Fig. 1c). Top, Pro/Anti signed d' and Choice signed d' as a function of time for each neuron. **c**, Matrix of Choice selectivity for the SC population. Neurons are sorted as in panel a. The absolute values of the d' shown in panel a and panel c are combined in Fig. 1d. **d**, **e**, **f**, same as panels a, b, c for the PFC population.

Extended Data Figure 4

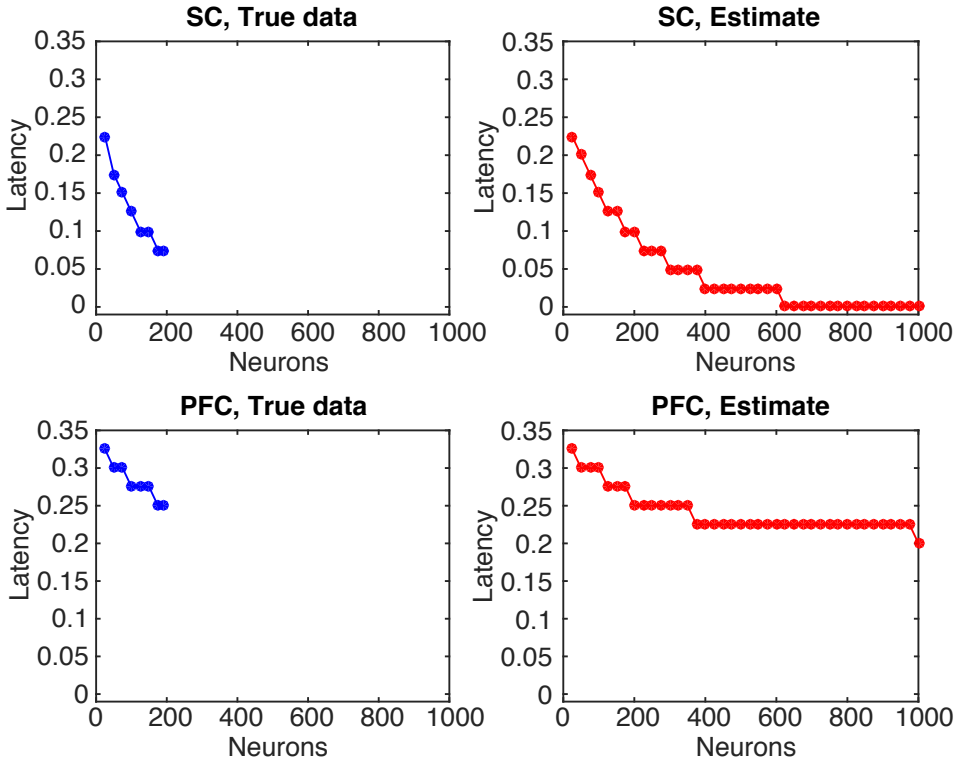
a **Result of removing spikes in SC to match the total number of PFC spikes**



b **SC performances can be reached by PFC by increasing the number of neurons to 1,000**

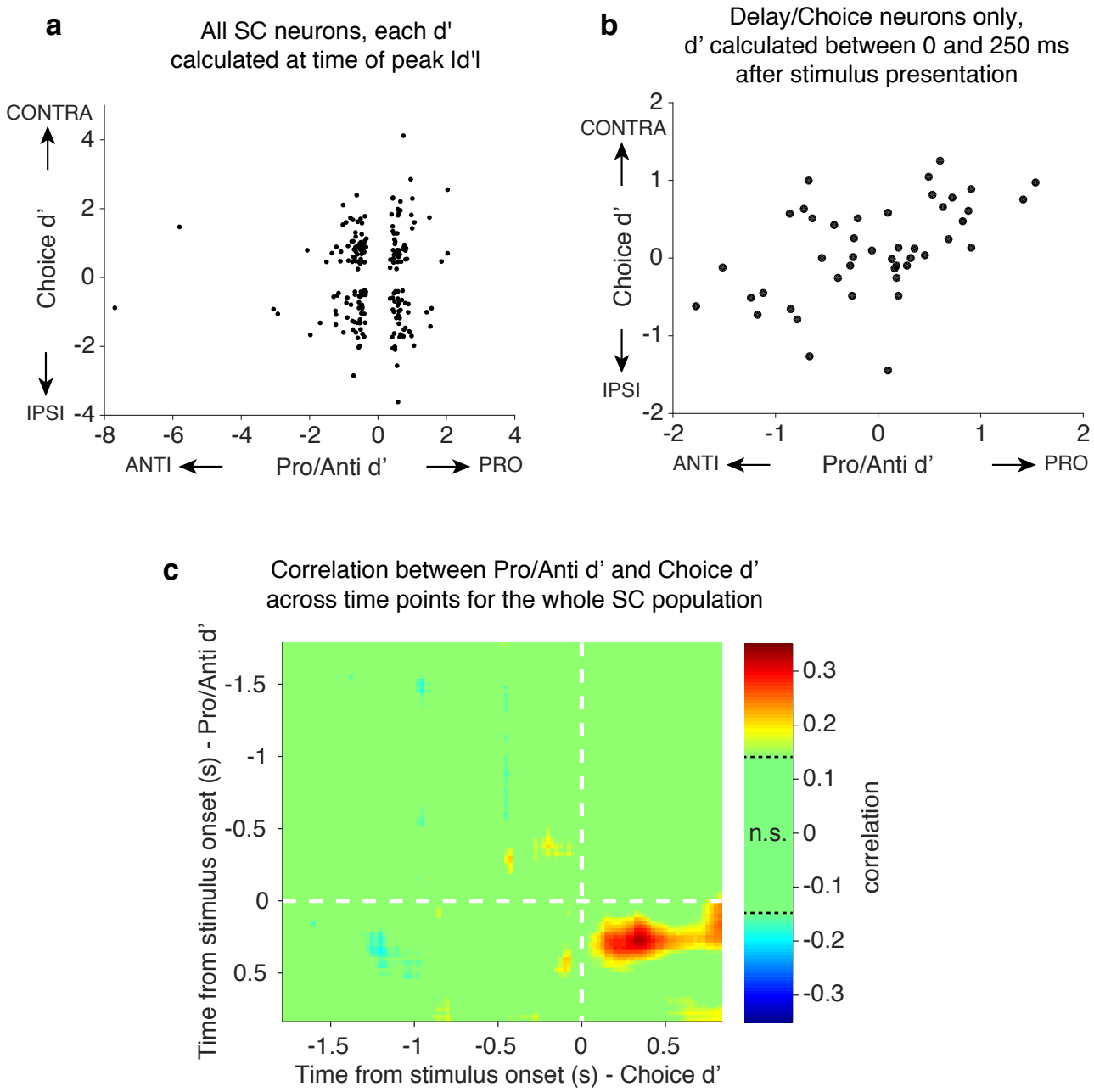


c **Latency in PFC cannot be reduced to SC levels by increasing number of neurons**



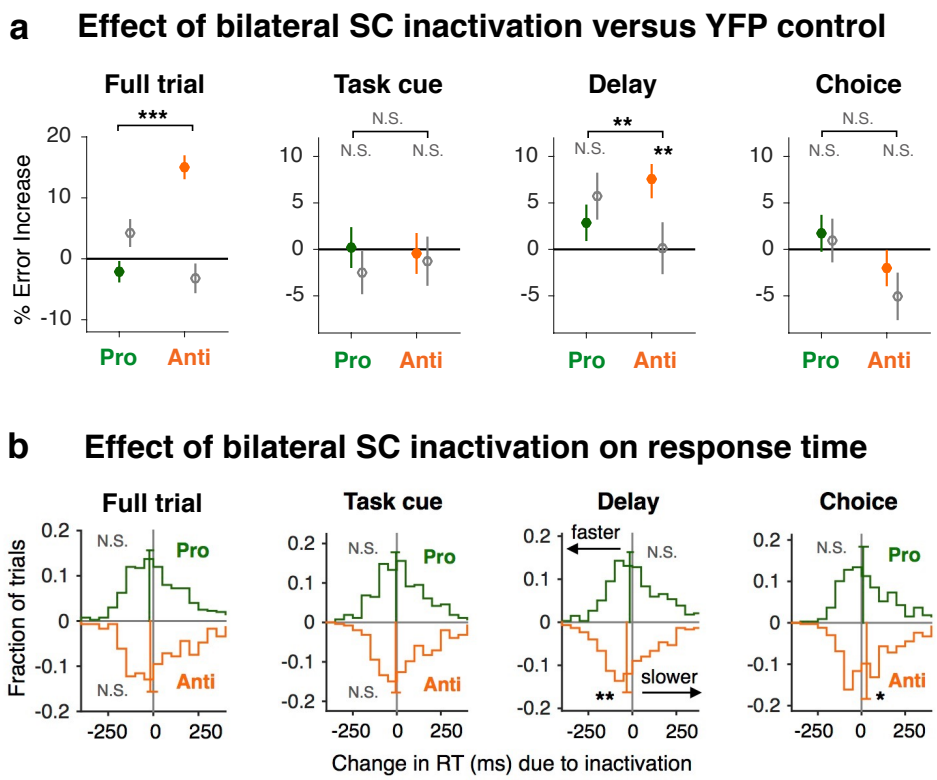
Extended Data Figure 4 | Controls for the comparison of population performances in SC and PFC. a, Pro/Anti (red, left) and Choice (blue, right) classification performances in SC (solid line) , PFC (dashed line), and SC after matching the average firing rate by randomly removing spikes (dash-dot line). Pro/Anti performances in SC are still significantly higher than PFC after matching firing rates ($p<0.05$). Latency of the rise in choice classification is still shorter in SC after matching firing rates ($p<0.01$). **b**, Estimated Pro/Anti classification performances for different number of neurons in SC (red) and PFC (black) (see Methods). Performances of 193 SC neurons can be matched by approximately 1000 PFC neurons **c**, Estimated latency of choice performances in PFC (blue, left) and in SC (red, right) for different numbers of neurons (see Methods). SC latency cannot be matched by PFC even when considering a population of 1000 neurons.

Extended Data Figure 5



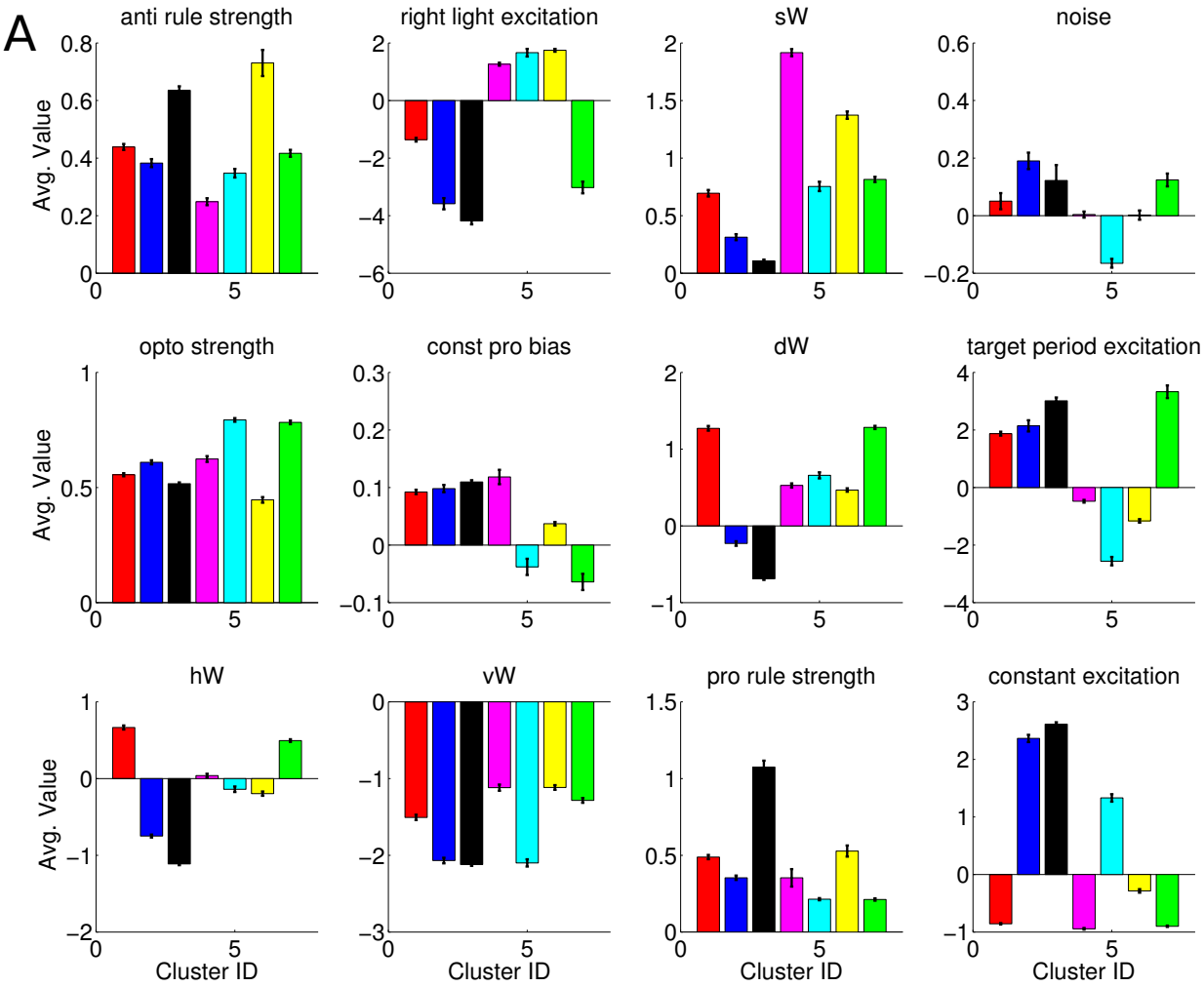
Extended Data Figure 5 | Relationship between task (Pro/Anti) and choice (Contra/Ipsi) d' across the SC population. a, For each SC neuron, the signed Pro/Anti d' computed at the time of peak Pro/Anti selectivity was plotted against the signed Choice d' computed at the time of peak Choice selectivity. No correlation is observed (r = 0.06). **b**, For SC Delay/Choice neurons, the signed Pro/Anti d' was plotted against the signed Choice d', both computed within the first time bin after stimulus appearance (0-250 ms). The two are significantly correlated (r = 0.52), due to the prevalence of Pro/Contra and Anti/Ipsi units. **c,** Correlation between Pro/Anti d' and Choice d' for the whole SC population computed at all time points. The correlation is significantly different than 0 only at times shortly after the appearance of the target stimulus.

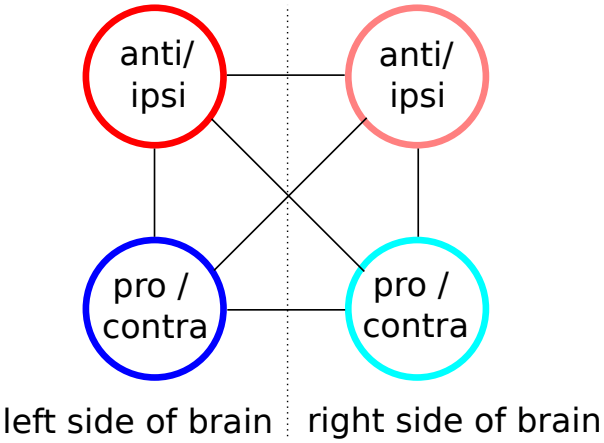
Extended Data Figure 6

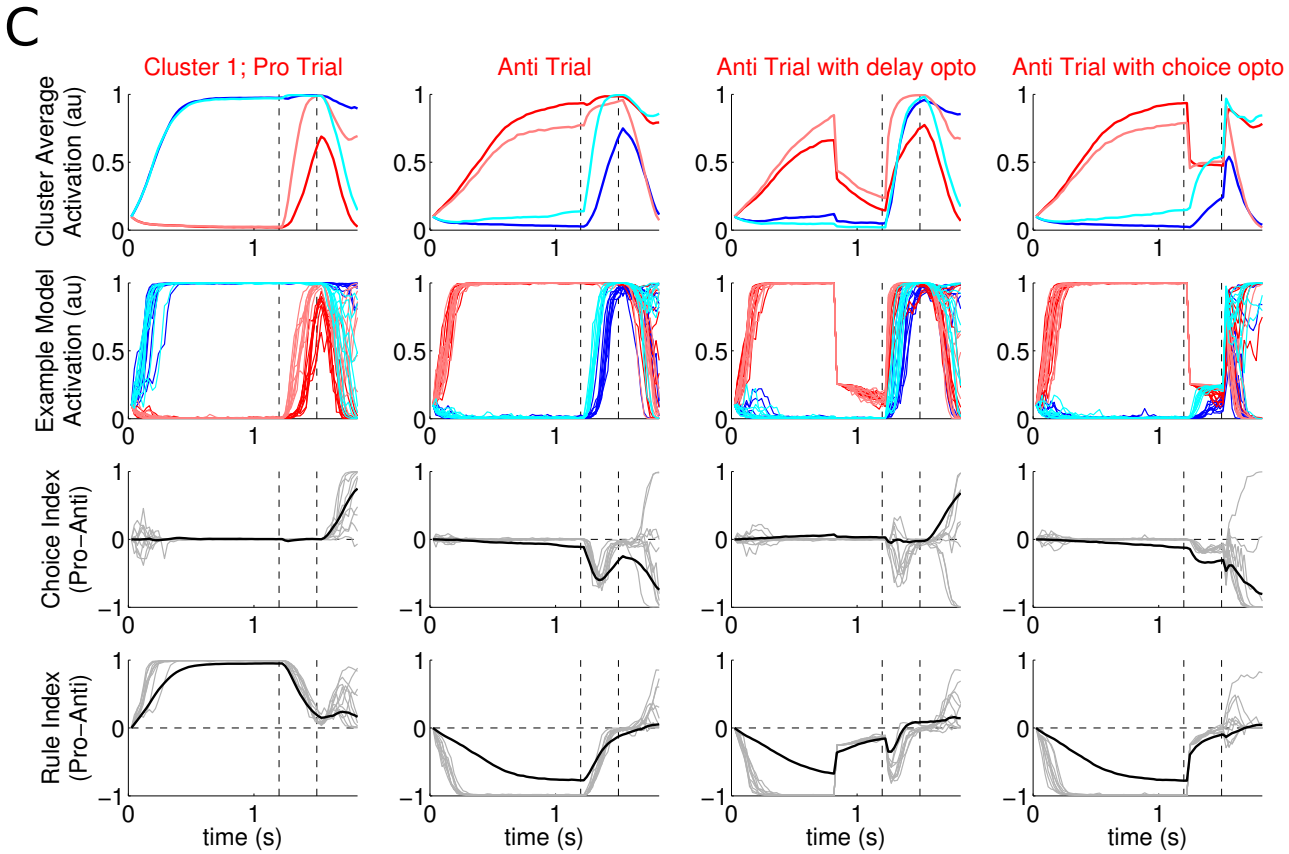


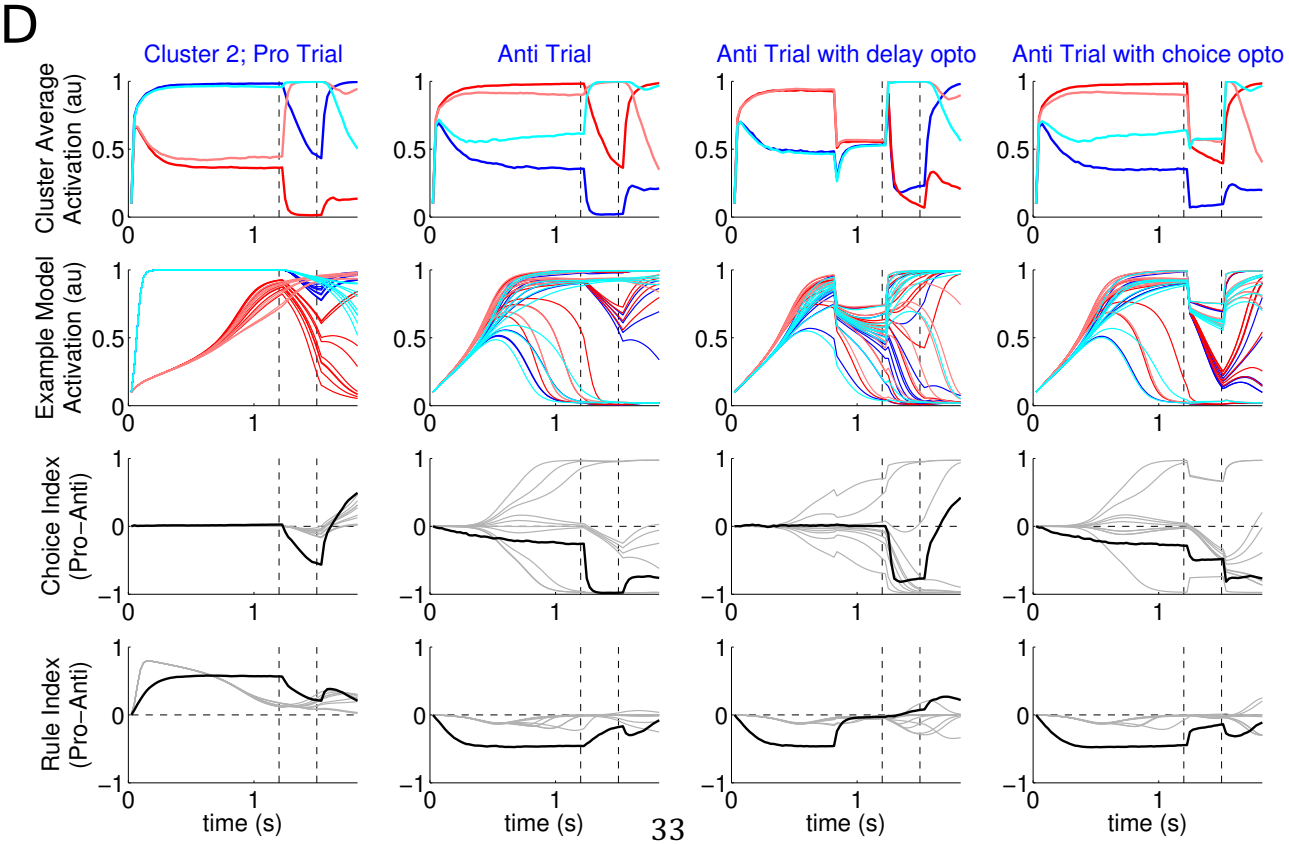
Extended Data Figure 6 | Effect of bilateral SC inactivation and YFP control. a, Effect of full-trial and sub-trial inactivations of bilateral SC on Pro (green) and Anti (orange) error rate (mean and s.e.m.) compared to YFP controls (gray). All paired statistics shown here are computed using a permutation test, shuffled 5000 times. **b**, Effect of full-trial and sub-trial inactivations of bilateral SC on response time (RT). For each behavioral session, a median RT on non-stimulated control trials is calculated and subtracted from the RTs on inactivation trials, and these normalized RT changes due to inactivation are plotted here. Each curve is normalized to have a total area of 1. Vertical bars show the median RT changes for correct Pro and Anti trials; 95% confidence intervals across trials for each trial type are indicated by horizontal bars. A shift to the right indicates slowing due to inactivation and a shift to the left indicates speeding. N.S. $p>0.05$; *$p<0.05$; **$p < 0.01$; ***$p < 0.001$.

Extended Data Figure 7

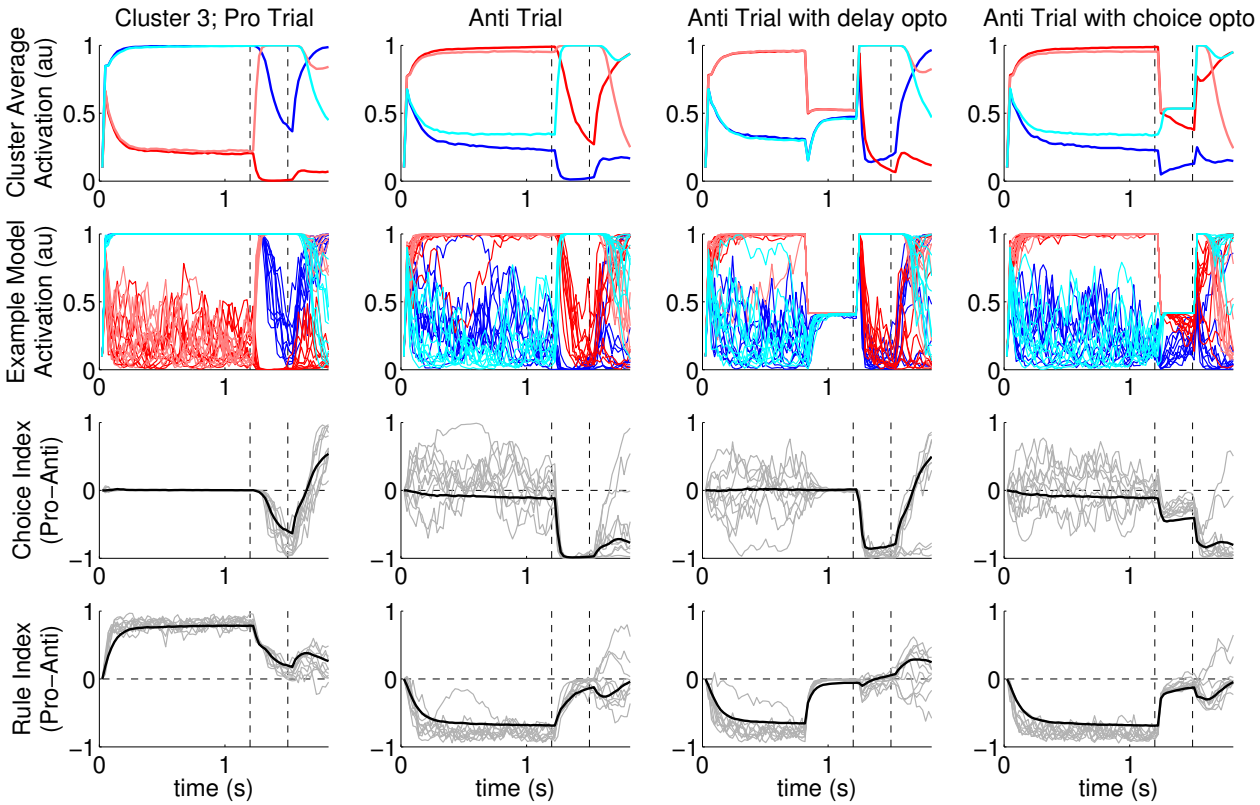




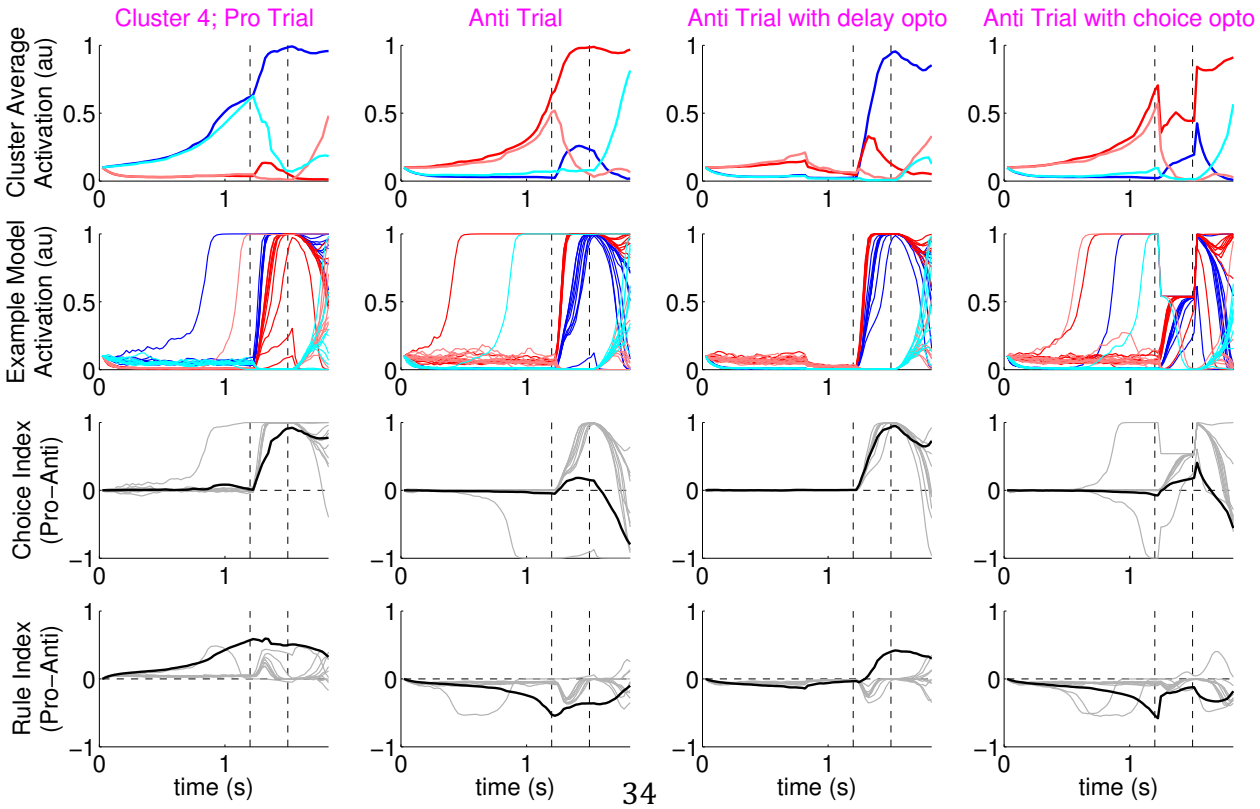

C
Cluster 1; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)


D
Cluster 2; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)

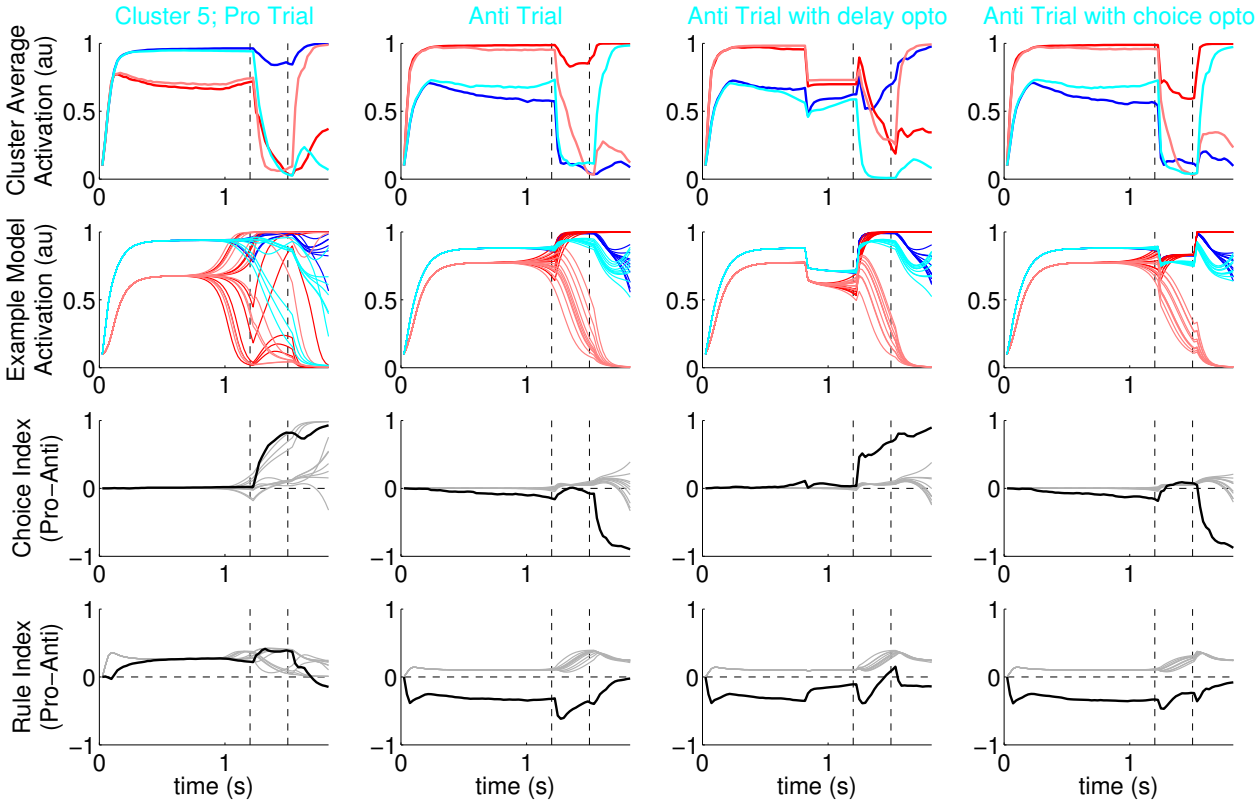
E


Cluster 3; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)

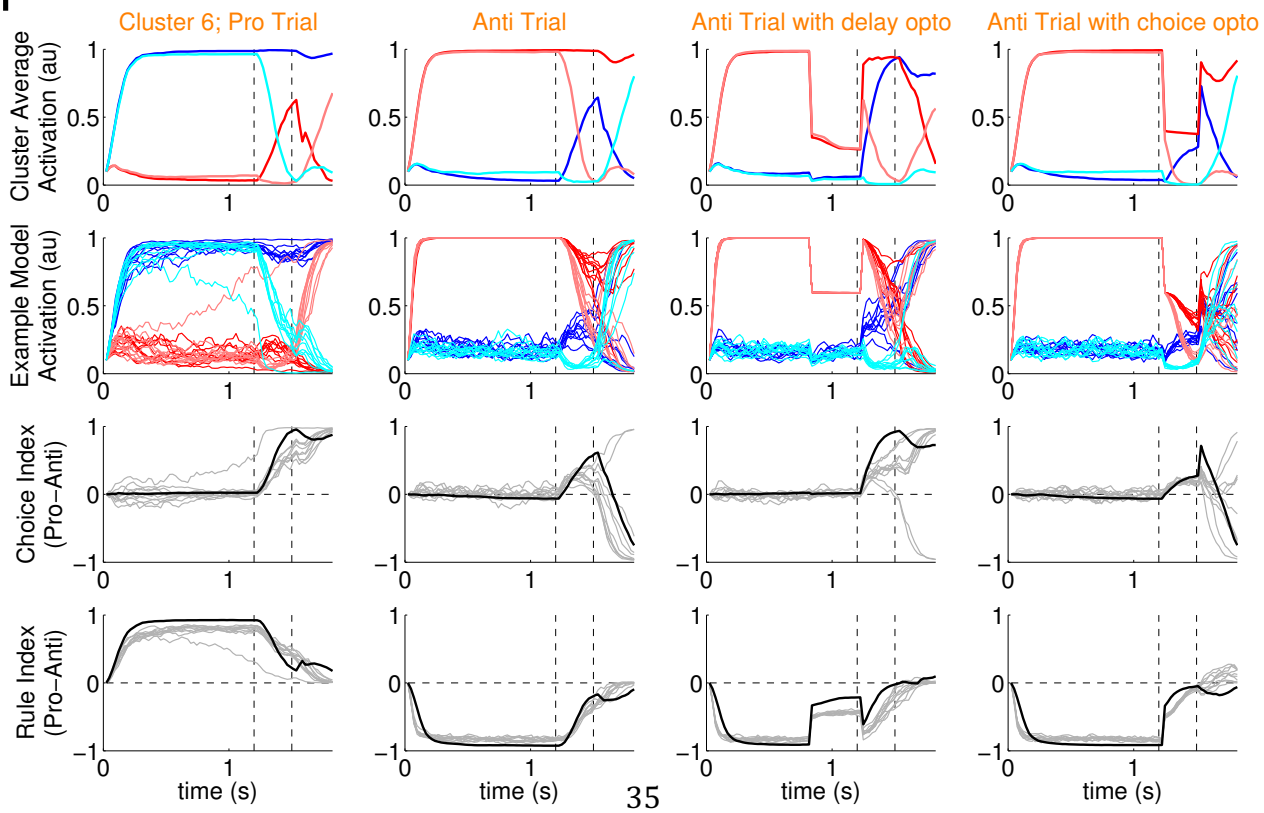
F

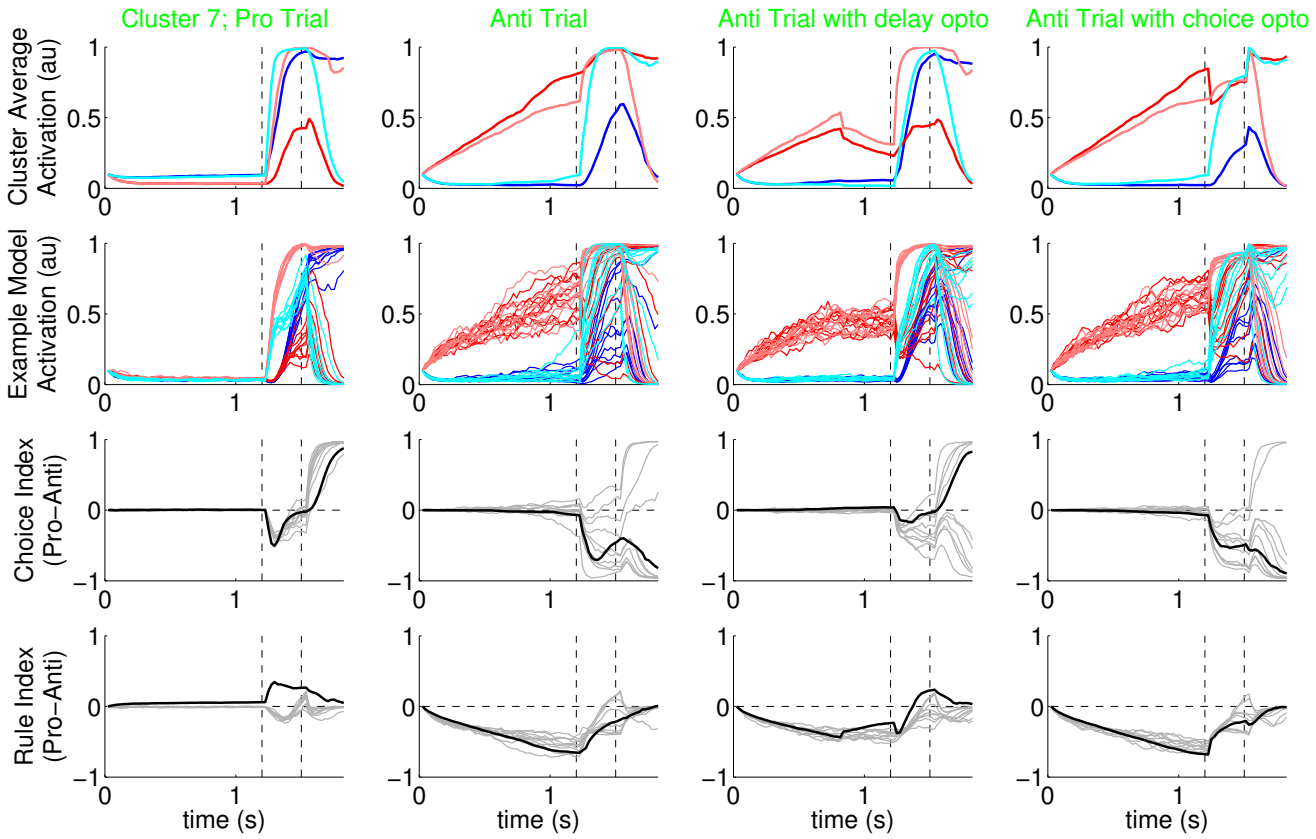

Cluster 4; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)

G


Cluster 5; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)

H


Cluster 6; Pro Trial
Anti Trial
Anti Trial with delay opto
Anti Trial with choice opto
Cluster Average Activation (au)
Example Model Activation (au)
Choice Index (Pro−Anti)
Rule Index (Pro−Anti)
time (s)



Extended Data Figure 7 | Parameters and dynamics of each cluster of model solutions. a, Mean and standard deviation of each parameter for each of the 7 clusters of model solutions as reported in the main text. Color scheme consistent with main figure. **b**, Schematic of model showing color scheme used for dynamics of each unit in panels c-i. **c-i,** For each cluster, the four columns are: Pro control trials, Anti control trials, Anti trials with delay period inactivation, Anti trials with choice period inactivation. Top row: average PSTHs across all model solutions in cluster. 2nd row: 10 example trials from one model solution picked at random from each cluster. 3rd row: choice encoding index is Pro_right - Pro_left. Black line is index for average PSTH, grey lines are index for example trials. 4th row: rule encoding index is 0.5*(Pro_right + Pro_left) - 0.5*(Anti_right + Anti_left).