Primate thalamic nuclei select abstract rules and shape prefrontal dynamics

Jessica M. Phillips[1,2]*, Mohsen Afrasiabi[1]^, Niranjan A. Kambi[1]^, Michelle J. Redinbaugh[1], Summer Steely[1], Emily R. Johnson[1], Xi Cheng[3,4], Maath Fayyad[1], Sounak Mohanta[1], Asia Carís[1], Charles B. Mikell[3], Sima Mofakham[3,4], Yuri B. Saalmann[1,2]*

[1]Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson St., Madison, WI 53706, United States
[2]Wisconsin National Primate Research Center, University of Wisconsin-Madison, 1202 Capitol Ct., Madison, WI 53715, United States
[3]Department of Neurosurgery, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY 11794, USA
[4]Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794, USA

^ denotes equal contribution
* Corresponding authors:
jphillips@primate.wisc.edu
Room 218, 1220 Capitol Court
Madison WI 53715, USA
saalmann@wisc.edu
518 Brogden Building, 1202 W Johnson St
Madison, WI 53706, USA

SUMMARY

Flexible behavior depends on abstract rules to generalize beyond specific instances, and outcome monitoring to adjust actions. Cortical circuits are posited to read out rules from high-dimensional representations of task-relevant variables in prefrontal cortex (PFC). We instead hypothesized that converging inputs from PFC, directly or via basal ganglia (BG), enable primate-specific thalamus to select rules. To test this, we simultaneously measured spiking activity across PFC and two connected thalamic nuclei of monkeys applying rules. Abstract rule information first appeared in the ventroanterior thalamus (VA) – the main thalamic hub between BG and PFC. The mediodorsal thalamus (MD) also represented rule information before PFC, which persisted after rule cues were removed, to help maintain activation of relevant posterior PFC cell ensembles. MD, a major recipient of midbrain dopamine input, was first to represent information about behavioral outcomes. This persisted after the trial (also in PFC). A PFC-BG-thalamus model reproduced key findings, and thalamic-lesion modeling disrupted PFC rule representations. These results suggest a revised view of the neural basis of flexible behavior in primates, featuring a central role for thalamus in selecting high-level cognitive information from PFC and implementing post-error behavioral adjustments, and of the functional organization of PFC along its anterior-posterior dimension.

INTRODUCTION

PFC neurons respond to different task variables (mixed selectivity[1]) and adapt to current task demands (adaptive coding[2]), to form a large, flexible workspace. Such a high-dimensional representation allows readout of task-relevant rules, to guide behavior or be fed back to PFC neurons to influence processing of upcoming task events for hierarchical cognitive control[1,3]. Although downstream cortical neurons have been posited to perform this readout, convergent PFC projections to the thalamus directly, or via the basal ganglia (BG), suit thalamic readout, which would be disseminated via divergent thalamo-PFC projections. This would enable strengthening of initially weak PFC signals, updating of PFC representations and adaptive coding.

The mediodorsal thalamus (MD) contains mostly "core" neurons innervated by cortical layer 6, which form local thalamo-cortical loops and have a driving influence on PFC middle layers[4,5] (Figure 1G). In mice, such loops reinforce PFC representations, important for working memory processes[6,7]. But mouse MD neurons show little/no rule information and sustain PFC representations via a nonspecific volley of activity[7]. This may be related to the reduced cognitive capacity and relatively low numbers of rule-selective neurons in mouse PFC (compared with primates). We predicted that primate MD would carry substantial rule information across delay periods. Another species difference is that primate MD receives major dopaminergic input[8], suggesting an additional outcome monitoring role.

Only found in primates[5,9], the ventroanterior thalamus, magnocellular compartment (VAmc), contains more "matrix" neurons innervated by cortical layer 5, which participate in long-range thalamo-cortical paths and have a modulatory influence in PFC superficial layers[4,5] (Figure 1G). Because VAmc receives major BG outflow from the substantia nigra pars reticulata (SNr) in addition to PFC inputs[10,11], we hypothesized that these multiple levels of convergence enable VAmc neurons to select task-relevant rules.

RESULTS

Diffusion MRI-guided electrophysiological recordings

Two monkeys learned the hierarchical rule task (HRT), requiring the sequential application of two rules (Figure 1A,B; Figure S1). It is hierarchical in the sense that the first, abstract rule (e.g., orientation) specifies a class of subordinate, more concrete rules (e.g., SW, N, SE), similar to tasks used with humans[12,13]. While monkeys performed the HRT, we used linear microelectrode arrays (LMAs) to simultaneously record from four cortical and subcortical regions of interest (ROIs; Figure 1C,D). For PFC recording sites, we targeted the principal sulcus (putative areas 46d/v) in two locations separated by approximately 7 mm in the anterior-posterior plane, thus yielding anterior PFC (aPFC) and posterior PFC (pPFC) sites in each monkey. We used diffusion MRI and probabilistic tractography to target locations in MD and VA with a high likelihood of connectivity with the PFC sites (Figure 1D, blue voxels). For VA, we additionally targeted voxels which had a high probability of connection with the SNr, indicating the likely location of VAmc (Figure 1D, A19.5 panel, blue voxels). Thus, our thalamic LMAs were situated in central MD (MDc, putative MDpc) and medial VA (VAm, putative VAmc).

We measured spiking activity in a total of 1094 neurons in four ROIs (229 in aPFC, 340 in pPFC, 196 in VAm, 329 in MDc; all included in whole-ROI analyses). In cortex, we used current source density[14] (CSD; Figure 1E) and unipolar local field potential (LFP) analyses[15,16] (Figure 1F) to designate electrode contacts and neurons to cortical layers (n = 147 superficial, n = 104 middle, n = 272 deep; all included in layer-specific analyses; remaining PFC cells could not be assigned to a particular layer due to partial laminar span in that recording session). Of the aPFC and pPFC neurons, we designated 65 and 82 to superficial, 54 and 50 to middle and 105 and 167 to deep layers, respectively.

Thalamus selects abstract rules

Many thalamic neurons showed robust, short-latency selectivity for the currently relevant abstract rule (Figure 2A-H; Figure S2A-F). We used two different methods to quantify selectivity of neurons. First, we used a relatively conservative nonparametric test to compare the number of spikes across conditions, which suggested a subset in each population was rule selective (aPFC, 2.6%, Figure S2G; pPFC, 10.3%, Figures 2I, S2H; VAm, 15.3%; MDc, 12.5%). Linear trend analysis leveraging a General Linear Model (GLM) and linear contrasts revealed that VA cells achieved selectivity at the shortest latency (100ms; $p<0.001$; $t>3.4$), followed by MD (150ms; $p<0.008$; $t>3.1$), and then pPFC (250ms; $p<0.01$; $t>2.8$). The overall latency followed the pattern: VA<MD<pPFC ($p<0.032$; $t>2.4$; Figure 2J). Second, using an abstract rule-selectivity index (SI)[17,18], we identified single neurons in each ROI as abstract-rule selective, and calculated the population average abstract rule SI timecourse for each ROI (Figure 2D,H). Results from both methods suggested that MDc and VAm populations carry just as much, if not more rule information compared to PFC populations.

To confirm this, we decoded the abstract rule from pseudopopulation spiking activity of all neurons for each ROI across all recording sessions (Figure 2K). Abstract rule information in thalamic populations was greater, and available earlier, than that from PFC populations. During abstract rule instruction (0-450 ms), VAm had the highest decoding accuracy ($t > 5.1$; $p < 0.0007$; N=100000; ANOVA) and shortest latency ($t > 3.4$; $p < 0.002$; N=100000; ANOVA, 115 ms), followed by MDc (magnitude, $t > 3.1$; $p < 0.006$; latency, $t > 2.88$; $p < 0.01$; N=100000; ANOVA, 135 ms) then pPFC (magnitude, $t > 6.2$; $p < 0.0001$; latency, 150 ms), while aPFC was not informative. This suggests that PFC neurons inherit robust rule selectivity from thalamic neurons; with BG input possibly contributing to the prominent early role of VAm. Consistent with this, in the pPFC, abstract rule information first appeared (latency 155 ms) in superficial ($t > 1.89$; $p < 0.02$; N= 100000; ANOVA) and deep ($t > 2.23$; $p <$

0.02) layers – both targeted by VAmc – followed by middle layers ($t > 1.45$; $p < 0.04$; latency 165 ms) – targeted by MDpc (Figure 2L).

The number of rule-selective neurons in pPFC increased across the first delay when monkeys had to keep the abstract rule in working memory (Figure 2M,N). Decoding the abstract rule during this delay (Figure 2K,O) suggests pPFC ($t > 2.45$; $p < 0.006$; N=100000; ANOVA) and MDc ($t > 1.8$; $p < 0.02$) carried more rule information than aPFC and VAm; with pPFC carrying more rule information than MDc ($t > 1.3$; $p < 0.03$). Superficial layers of the pPFC were most informative ($t > 4.76$; $p < 0.0003$; Figure 2L) and MDc maintained rule information across the delay (Figure 2O,P). As a control for possible differences in spiking variability between areas which could influence decoding accuracy, we calculated the mean-matched Fano Factor[19] and there was no significant difference between the areas (Figure S3). This suggests recurrent PFC-MDpc loop activation helps sustain the abstract rule once it has been established by transient VAmc activity.

Concrete rule selectivity first appears in PFC

Both PFC and thalamic neurons exhibited robust, short-latency concrete-rule selectivity (Figure 3A-D; Figure S4). The nonparametric tests indicated neurons in the pPFC (23.6%), VAm (8.2%) and MDc (12.2%) were selective for the concrete rule. This concrete-rule selectivity could arise first in the thalamus via convergent cortical inputs (similar to abstract rule), or it could arise first in PFC as the rule set bias has already been established in the PFC following abstract rule processing, enabling sensory inputs to effectively trigger PFC representation of the relevant concrete rule. We performed a population analysis of concrete rule SI in each ROI, to test whether concrete rule information in PFC and thalamus differs (Figure 3E,F). This suggested that pPFC neurons carry more concrete rule information than those of VAm and MDc; during cue presentation, MDc may carry more information than aPFC.

The pseudopopulation decoding analysis confirmed that the pPFC population carried concrete rule information earliest ($t > 3.15$; $p < 0.004$; N=100000; ANOVA; latency 105 ms) and most robustly ($t > 6.2$; $p < 0.00005$; N=100000; ANOVA) during both concrete rule cue presentation and the second working memory delay period. MDc was next most informative ($t > 4.87$; $p < 0.0003$; latency 145 ms) followed by VAm ($t > 3.02$; $p < 0.008$; latency 130 ms), and aPFC was not informative about the concrete rule (Figure 3G). Confusion matrices for each ROI demonstrated decoding of each of the six concrete rules, except in aPFC (Figure S5). In pPFC (Figure 3H), concrete-rule selectivity appeared first ($t > 5.78$; $p < 0.00004$; N=100000; ANOVA; latency 105 ms) and most robustly ($t > 3.008$; $p < 0.0008$; N=100000; ANOVA) in the deep layers, followed by middle layers (latency 125 ms), then superficial layers (latency 160 ms). This is consistent with an intracortical effect first giving rise to PFC concrete-rule selectivity, after which core thalamic neurons (particularly in MDpc) projecting to PFC middle layers help sustain concrete rule representations. Thalamic matrix cells projecting to PFC superficial layers seem to play a minor role in generating concrete-rule selectivity.

Strong directional selectivity in pPFC supported by MD

Previous studies have shown directional target/saccade information in PFC and thalamus[20]. The nonparametric tests showed single neurons in each ROI exhibited direction selectivity: aPFC, pre 4.8%, peri 5.4%, post 7.9%; pPFC, pre 30%, peri 18.7%, post 22.9%; VAm, pre 15.9%, peri 14.7%, post 11.6%; MDc, pre 20.4%, peri 15.7%, post 11.3%. We calculated the directional SI timecourse for single neurons

found to be selective during pre-, peri- and post-saccade windows (Figure 4A,B,D, Figure S6) for population direction SI analyses, which suggested that pPFC carried more directional information than both thalamic nuclei prior to, during and following saccades (Figure 4C,F). Both thalamic nuclei carried more directional information than aPFC during at least part of the pre-saccade window. During the peri-saccade window, directional information appeared stable across all ROIs except VAm, which had less directional information than aPFC. After saccade onset, pPFC directional information appears to peak, then decrease along with MDc, while directional information was similar for aPFC and VAm.

The pseudopopulation decoding analysis of directional selectivity aligned with our univariate results. pPFC carried the most robust directional information followed by MDc, VAm then aPFC ($t > 2.99$; $p < 0.004$; N=100000; ANOVA; Figure 4E). Confusion matrices for each ROI confirmed decoding of each direction (Figure S7). The robust directional information in pPFC reflects greatest information carried by its deep layers, followed by middle then superficial layers ($t > 3.63$; $p < 0.001$; N=100000; ANOVA; Figure 4H). Similarly, in aPFC, directional information was greater in deep, middle, then superficial layers ($t > 3.4$; $p < 0.002$; N=100000; ANOVA). These patterns are consistent with PFC output from deep layers driving directional information in their targets before, during and following saccades.

Outcome information processed in MD-PFC circuits

One important aspect of cognitive control is monitoring whether chosen actions achieve desired outcomes. Many neurons in each ROI had activity modulated by trial outcome (Figure 5A-F, Figure S8). The nonparametric tests yielded outcome selective neurons in every ROI (aPFC, 40.1%; pPFC, 25.4%; VAm, 21.0%; MDc, 52.0%). Some neurons showed greater activity following errors compared to correct trials, while others showed the opposite pattern, and yet others showed a more complex dynamic modulation. Some neurons showed sustained modulation by trial outcome, persisting across the intertrial interval. Such a pattern may indicate an influence on the allocation of cognitive resources on the following trial or updating of reward expectations to optimize decision-making.

The pseudopopulation decoding analysis showed that each ROI carried information on trial outcome, but MDc expressed this information earliest, followed by VAm, aPFC, then pPFC ($t > 0.67$; $p < 0.03$; N=100000; ANOVA; Figure 5G). During the first 1000 ms post-saccade, MDc carried most information on outcome, followed by aPFC, pPFC, then VAm ($t > 1.9$; $p < 0.003$; N=100000; ANOVA). Afterward, MDc continued to carry the most information, while aPFC, pPFC and VAm populations carried similar outcome information ($t > 2.93$; $p < 0.0001$; N=100000; ANOVA). However, there were differences in decoding outcome by PFC layer. In aPFC and pPFC, outcome information emerged earliest ($t > 1.32$; $p < 0.01$ and $t > 1.2$; $p < 0.03$, respectively) and was strongest ($t > 2.1$; $p < 0.007$ and $t > 2.54$; $p < 0.005$, respectively) in deep layers, comparable to MDc. In pPFC, middle layers carried greater outcome information than superficial layers. This is consistent with recurrent thalamo-PFC loop activation involving thalamic core neurons sustaining outcome information. The latency results suggest that this information may originate in MDpc[21].

PFC-BG-thalamus model reproduces in vivo activity

To characterize the mechanism behind selection and maintenance of abstract rules, we built a layer-specific PFC-BG-thalamic model comprising leaky integrate-and-fire neurons, to process the

abstract rule cue and working memory information across the subsequent delay period in the HRT (Figure 6A). We decoded the abstract rule from the spiking activity of modeled neurons for each ROI. Initially after cue onset, superficial layer PFC neurons showed sensory-selective responses (raster responses to one of the four cues only), with inhibition during non-preferred cue presentation (Figure S9A-D). Deep layer PFC neurons showed relatively less sensory selectivity, due to the convergence across PFC layers (Figure S9E-H and S10). However, abstract rule selectivity first emerged in the thalamus, due to the extra cortico-thalamic convergence (Figures 6B, S9K-N). VAmc had the shortest latency to rule-selective responses (50 ms; t=4.2, p=0.0006), then MDpc (150 ms; t=3.81, p=0.001). The cue-induced BG disinhibition of VAmc produced a facilitating time window for VAmc neuron firing (Figure S9K, M). Abstract rule information was then sent to PFC (200 ms latency; t=3.22, p=0.002), and maintained across the delay period in PFC-MDpc loops (Figure S9I, J, L, N).

Thalamic lesion modeling disrupted PFC rule representations

We next modeled lesions of VAmc or MDpc by disconnecting them from the rest of the network, to test influences on abstract rule processing in other circuit nodes. Modeled VAmc lesions markedly reduced rule selectivity at all time points after cue onset (although latencies did not differ: MD 150 ms; t=2.22, p=0.002; PFC 200 ms; t=2.88, p=0.005). Decoding accuracy dropped about 20% during cue presentation for both MDpc and PFC and, during the delay period, decoding declined further for MDpc and was at chance levels for PFC (Figure 6D). In comparison, modeled MDpc lesions did not affect rule representations in PFC and VAmc during cue presentation (VA 50 ms; t=4.2, p=0.0006; PFC 150 ms; t=3.65, p=0.001). However, decoding was at chance levels for VAmc and PFC during the delay period (Figure 6C). This supports the mechanism where VAmc driving projections transmit abstract rule information to PFC; and VAmc's modulatory projections regulate cortico-cortical connections, allowing recruitment of additional PFC neurons to the rule representation, necessary to bridge the cue and delay periods. MDpc driving input to PFC yields synchronized activation of PFC neurons, which recruits more MDpc neurons to support rule information maintenance across the delay. Finally, we created model regimes with stronger thalamo-cortical weights (taking the model in the opposite direction of the lesions) and found inappropriate activation of rule representations outside trials (which one might speculate could lead to inappropriate behaviors in psychopathology). In model regimes where thalamo-cortical weights were relatively weak (our baseline regime), both VAmc and MDpc were needed to select and maintain abstract rules, highlighting the importance of matrix-like neurons to promote and shape PFC processing modes, in addition to core-like neurons supporting ongoing operations.

Overall, we have shown that robust abstract rule information is first available in the thalamus, especially matrix-heavy VAm, which is then issued to PFC superficial layers, establishing a cortical bias for the relevant rule set. This is supported by convergent patterns of cortical inputs to thalamus enabling abstract rule selection, and divergent outputs to PFC disseminating the rule[4]. Once this occurs, circuits involving MD core neurons and PFC may provide the drive necessary to maintain the bias, facilitating optimal future computations, poising the system to efficiently transform subsequent inputs to rules and responses once cognitive control demands lessen, permitting more concrete control over behavior. Our model suggests that the thalamus provides a "sweet spot" for cortical excitability and connectivity, in which cortical neurons can be flexibly recruited in different tasks and, in each task, be noise-robust, maintaining signal fidelity.

It has been reported that VA neurons encode stable stimulus values[22], and that MD neural activity is predictive of behavioral outcome[23]. Here we show that thalamus represents outcome information early,

and before aPFC, suggesting thalamus may inform cortex and a necessary revision to current frameworks emphasizing orbital and medial frontal contributions[4]. This fits the anatomy showing primate MD, along with PFC deep layers, receives dense innervation from midbrain dopaminergic centers[8,24], perhaps associated with unique dopaminergic mechanisms (compared to those in BG). The time-course of outcome information is consistent with a spread from anterior-to-posterior PFC, either directly or via thalamus. MD and PFC represented the preceding response and its outcome, needed for updating reward expectations to optimize decision-making[25].

Based on influential models positing that the anterior-to-posterior dimension of PFC is functionally organized according to abstraction level[26,27], we expected aPFC to strongly encode abstract rules, relative to pPFC. In contrast, we found aPFC carries little information about specific task events, and instead appears to play a prominent role in monitoring outcomes, consistent with aPFC supporting metacognition[28]. pPFC, on the other hand, represented both abstract and concrete rules, inconsistent with a PFC topography based on policy abstraction level. However, the aPFC role in outcome monitoring at longer timescales suggests there may be a topography based on temporal abstraction[27,29]. Because VAm and MDc selected and maintained abstract rules, PFC's functional organization needs to incorporate thalamic loops, which select abstract information from high-dimensional cortical representations, then update PFC representations to shape task-relevant manifolds.

METHODS

Animals

We acquired data from two male rhesus monkeys (*Macaca mulatta*, age 7 and 9 years, weighing 9.5 and 10 kg, respectively). The daily needs of the animals were met by experimenters and husbandry staff at the Wisconsin National Primate Research Center (WNPRC), where the animals were housed. The health of the animals was monitored by WNPRC veterinary staff. All procedures were approved by the University of Wisconsin-Madison Institutional Animal Care and Use Committee, and conforms to the National Institutes of Health Guide for the Care and Use of Laboratory Animals

Neuroimaging data acquisition

Prior to implant surgery, we acquired a high-resolution structural brain image to delineate thalamic and cortical regions of interest (ROIs). We also acquired diffusion-weighted MRI data, to identify interconnected network sites and target subregions of interest inside thalamic nuclei. Following craniotomy, we acquired additional structural brain images with electrodes *in situ* to confirm their positioning.

We performed diffusion-weighted and structural imaging on each monkey using the GE MR750 3T scanner (GE Healthcare, Waukesha, WI). Monkeys were pre-medicated with ketamine (up to 20 mg/kg body weight) and atropine sulfate (0.03-0.06 mg/kg) prior to intubation. Subsequently, isoflurane was administered (1-3% on ~1 L/min O_2 flow) with a semi-open breathing circuit and spontaneous respiration, to maintain general anesthesia. We monitored the monkeys' vital signs (expired carbon dioxide, respiration rate, oxygen saturation, pulse rate, temperature) using an MR-compatible pulse oximeter and rectal thermometer.

For the three-dimensional T1-weighted structural volumes, we used an inversion-recovery prepared gradient echo sequence with the following parameters: FOV = 128 mm^2, matrix = 256 x 256, number of slices = 166; 0.5 mm isotropic; TR = 9.68 ms; TE = 4.192 ms; flip angle = 12°; and inversion time (TI) = 450 ms. We collected 10 T1-weighted structural volumes and calculated the average image for each monkey using the FMRIB Software Library (FSL)[30]. During diffusion scan sessions, this protocol was used to acquire an additional pair of structural volumes for co-registration purposes. Finally, to localize electrodes, we averaged two structural images with electrodes *in situ*.

We used 2D echo-planar imaging (EPI) with a pulse sequence spin echo gradient pulse[31] with parameters: field of view 128 mm x 128 mm, resolution 1.0 mm x 1.0 mm; 80 x 1.0 mm coronal slices; repetition time = 11 000 ms; echo time = 77.5 ms; and parallel imaging (ASSET) factor of 2, to reduce the echo-spacing and thus image distortions. The acquisition used partial Fourier encoding of 0.0625 in the phase encoding direction to reduce the echo time and the resulting images were reconstructed using homodyne processing.

We used two protocols for the acquisition of the diffusion MRI data[21]. For one monkey, we used protocol A, which involved positioning the monkey in the 16-channel receive-only head coil (MRI instruments) using fitted cushions, and acquiring 8 b=0 images and 60 b=1000 s/mm^2 images. The entire acquisition was repeated 13 times. We also acquired a B0 fieldmap using a three-dimensional gradient-echo sequence with three echoes and the iterative decomposition of water and fat with echo asymmetry and least-squares estimation (IDEAL) method[32], which was used for geometric unwarping of diffusion-weighted data.

For the second monkey, we used protocol B, which involved positioning the monkey in the 8-channel receive-only head coil (Clinical MR Solutions) using a customized MRI-compatible stereotaxic apparatus. The acquisition protocol was similar for this monkey, but with 10 b=0 images and 120 diffusion directions, and with a second set of b=0 images acquired using reversed phase encoding. The entire acquisition was repeated 10 times. The inclusion of the second set of reverse phase encode b=0 images enabled the use of newer utilities for distortion correction (the two acquisition protocols and preprocessing pipelines produced qualitatively similar results, see[21]).

Probabilistic tractography on diffusion-weighted neuroimaging data

We used FSL to preprocess the diffusion MRI data[33,34], and to calculate probabilistic connectivity maps using probabilistic tractography. Based on these connectivity maps, we could isolate voxels in the mediodorsal thalamic nucleus (MD) with high probability of connection to the principal sulcus (PS) in the dorsolateral prefrontal cortex (dlPFC), and in VA with high probability of connection to the substantia nigra pars reticulata (SNr), which anatomists have used to define VAmc[35]. VAmc has preferential connections with the entire PFC[4,36-38]. Thus, this approach enabled targeting of electrodes to interconnected sites in prefrontal thalamocortical networks based on each monkeys' specific anatomy.

Preprocessing: Correction for motion, eddy currents and susceptibility distortions

For the dataset acquired with protocol A, we used the FSL utility Eddy_Correct to correct for motion (all EPIs) and eddy currents (diffusion weighted). We geometrically unwarped EPIs using the fieldmap and magnitude images acquired in the same session, to correct for susceptibility distortions[39]. In detail, to obtain a transformation matrix between fieldmap space and diffusion space, we manually skull-stripped the magnitude image. This volume was then forward-warped according to the fieldmap, using FSL's

Utility for Geometrically Unwarping EPIs (FUGUE), and registered (affine with 12 degrees of freedom (DOF)[40]) to an averaged, skull-stripped non-diffusion reference volume. We then applied the corresponding transformation matrix to the fieldmap image (scaled to rad/s and regularized by a 2-mm 3D Gaussian kernel), aligning it with the non-diffusion reference volume, so that it could be used to unwarp the EPIs with the FUGUE utility. Next, we skull-stripped the T1-weighted structural brain image and co-registered it to the averaged, skull-stripped, corrected non-DWI reference volume (12 DOF) to produce the transformation matrix between the two spaces. The corrected diffusion and b=0 volumes were then averaged across scan repetitions to produce a single set of 68 volumes (8 b=0 and 60 b=1000) for further processing.

For the dataset acquired with protocol B, we used FSL's Topup utility[33] to estimate the field distortions caused by susceptibility artifacts. In detail, we registered all b=0 images acquired with both phase encode directions to create a single pair of images with higher signal to noise ratio for submission to Topup. This pair of distorted images was used to estimate the field, and the two images were combined into a single corrected volume. We then submitted the Topup output, along with the entire concatenated dataset, to FSL's Eddy utility. Eddy uses the Topup field estimates, and corrects for motion, susceptibility distortions and eddy current distortions simultaneously[41]. A new set of direction vectors is output by Eddy, which accounts for any motion correction, so these data were not averaged at any stage of processing or analyses. We also co-registered structural space to diffusion space, as described above.

Estimation of diffusion parameters and probabilistic tractography

For the probabilistic tractography analyses, cortical and subcortical ROIs were manually delineated in each monkey using the high quality T1-weighted structural brain image and in consultation with stereotaxic atlases[42-44]. We applied the transformation matrix (from the co-registration of the structural image to the reference non-diffusion image) to the ROI masks for probabilistic diffusion tractography (PDT) analyses.

We performed the PDT analyses using FSL's Diffusion Toolkit (FDT). The algorithm modeled two fiber populations per voxel[45], which is suitable for the complex fiber architecture of the thalamus[46-48]. The posterior probability distributions of fiber direction at each voxel were calculated, for each monkey[46,49]. The anatomical plausibility of paths through thalamic and cortical ROIs was assessed as described in[21].

To identify MD voxels with a high probability of connection with the dlPFC (area 46, which contains our aPFC and pPFC recording sites), and VA voxels with a high probability of connection with the SNr, we performed a PDT analysis to estimate paths passing through any voxel in the thalamic ROI, and the probability that such paths will pass through a voxel in the non-thalamic ROIs, by running "multiple mask" tractography. Five thousand streamline paths were created from each seed voxel in the thalamic masks by drawing samples from their posterior probability distribution (of directions/angles) associated with the path's current voxel, and then adding the step to the end of the path based on the drawn angle (0.25 mm step length with maximum of 4000 steps). The sampling was repeated, with paths beginning in seed voxels of the non-thalamic ROI. The proportion of these samples (streamlines) passing through each voxel indicates the probability of connection to it. We used an exclusion mask to isolate ipsilateral connections and to eliminate anatomically implausible paths (see[21] for details). We used these paths to isolate "projection zones" in MD (from dlPFC) and VA (from dlPFC and SNr) using the method described in[21]. The locations of these projection zones were then used to guide planning for the craniotomy surgeries.

Surgery

Anesthesia was induced using ketamine (up to 20 mg/kg, i.m.) and general anesthesia maintained with isoflurane (1-2%) during aseptic surgical procedures. We used 12 ceramic skull screws with dental acrylic to affix cranial implants containing a head holder and craniotomy placeholder closed off with a custom plastic recording chamber. The cranial implant enabled behavioral training using video eye tracking. Once monkeys were fully trained on the hierarchical rule task (HRT), a second procedure was performed to place four 2.5 mm craniotomies in frontal and parietal bones. The craniotomies provided access to two cortical ROIs and two thalamic ROIs in the right hemisphere (aPFC, pPFC, MDc and VAm). Coordinates for craniotomies were calculated using the high-quality T1-weighted structural images acquired prior to the surgeries and the position of the projection zones estimated in the PDT analyses of diffusion-weighted data. We fitted each craniotomy with a conical plastic guide tube filled with bone wax, through which linear microelectrode arrays (LMAs) were advanced[48]. The guide tubes were prefabricated using a model of the skull based on the T1-weighted structural images. We also inserted a titanium skull screw positioned inside the recording chamber to serve as a reference.

Behavioral task and stimuli

Our ability to flexibly adapt behavior according to current goals and context is known as cognitive control. It is considered hierarchical, as goals can be selected at different levels of abstraction. An important way to implement cognitive control is to apply rules. To investigate thalamic contributions to hierarchical cognitive control and PFC processing, we designed an HRT. The task required monkeys to selectively process (based on initial abstract rule instruction cue) and map responses from (based on subsequent concrete rule instruction cue) one of two stimulus dimensions (while ignoring the other irrelevant stimulus dimension). Thus, each trial involved the presentation of a colored "abstract rule" cue, which informed the animal about the relevant dimension of the forthcoming white "concrete rule" cue (Figure 1). A red circle or blue triangle instructed monkeys to follow the abstract shape rule; whereas a green circle or yellow triangle instructed monkeys to follow the abstract orientation rule. The two stimuli to cue each abstract rule served as controls for purely visual responses in neurons. The monkeys had to keep this abstract rule information in mind over a brief delay. Importantly, the abstract rule cue did not specify a behavioral response, but one of two possible rule sets, and was thus a more abstract task parameter than the concrete rule. The concrete rule cue involved presentation of one of three possible shapes (bowtie, rectangle, oval), positioned according to one of three possible orientations (southwest, north, southeast). Each concrete rule cue shape and orientation mapped a response rule onto one of three targets, depending on the currently relevant abstract rule. Thus, monkeys could make their decision and prepare their response once they had been presented with the concrete rule cue. However, they had to maintain fixation over a second variable delay period until the fixation cross was removed and three targets were presented. There were nine possible permutations of the concrete cue (Figure 1A) and 36 experimental conditions; but at the abstract cognitive level, correct trials could be collapsed into six concrete rule categories (3 shape rules and 3 orientation rules).

The training stages involved focusing on one of the two abstract rule sets, to map a response to either shape or orientation (with the other stimulus dimension fixed, so that they could initially ignore it). The animals learned the first rule set by operant and classical conditioning (Figure S1, top row, "Bowtie only" and "North only"). Initially, there was no memory delay. Once the rule set (e.g., orientation rules) was learned on one stimulus feature (e.g., bowtie), then a second stimulus feature was included in the blocks of trials (e.g., ovals). The monkeys learned the stimulus-response mapping more quickly once they had already learned the rules on one stimulus feature, and thus showed generalization (Figure S1, middle row, "Bowtie + Oval"; "North + Southeast"), a hallmark feature defining abstract rule use. The third stimulus feature (e.g., rectangles) was learned the most rapidly (Figure S1, bottom row). The monkeys were then trained to perform the second abstract rule set (ignoring the first important dimension, e.g., orientation, and now attending to the second dimension, e.g., shape). We observed evidence of behavioral

generalization once again (Figure S1). Once both abstract rule sets had been learnt, the monkeys were exposed to both rule sets in the same behavioral training sessions and needed to use the instruction cue to select the abstract rule (which was not necessarily required in prior training sessions). Once the monkeys were able to switch between the two abstract rule sets across blocks within the same session, we then included the two abstract rule sets in the same block, randomly interleaved. The last step in training was to build in the memory delays. Monkeys overall performed at ≥ 65% accuracy (33% being chance performance) for our physiological recordings.

During each session, monkeys sat upright in a primate chair (Crist Instruments, Hagerstown MD) with their heads immobilized using a head post/head holder system (Neuronitek, London, Canada) and their eyes positioned 57 cm from a 19 inch CRT monitor (100 Hz refresh rate). The monkeys performed the HRT to earn juice rewards, indicating their target choices using eye movements. Eye position was monitored using a video eye tracker (SMI, Berlin, Germany). Behavioral task events, monitoring of eye position and feedback delivery were controlled by custom-written software in Presentation (Neurobehavioral Systems, Albany CA). Rewards were delivered via a TTL pulse issued to an infusion pump system (Harvard Apparatus, Holliston MA). The PC running Presentation issued TTL pulses to the PC running the neural data acquisition system (Plexon, Dallas TX), to store behavioral and task event information, together with neural data. Recording sessions were typically approximately 2 hours in duration and monkeys performed between 351-1047 trials during a recording session (736 trials on average).

Trials were offered to the monkeys via presentation of a white fixation cross (1.72° x 1.72°) at the center of the screen on an otherwise gray background. Monkeys initiated a trial by foveating the fixation cross, which was maintained on the screen for the duration of the trial until target presentation, to anchor the gaze across cue presentation and delay periods. Eye position was monitored within a 5° x 5° window, and if the gaze position exited the window prior to target presentation, the trial was aborted. After a variable preparatory period (500-1100 ms), the abstract rule cue was presented (either a 2.6° diameter circle or a triangle of 3.4° maximum width, duration of 450 ms, Figure 1A,B). This was followed by a variable (500-900 ms) delay period, where the monkey was required to hold the cue information in working memory. Next, the concrete rule cue was presented (either a rectangle, oval or bowtie which were framed by a rectangle of dimensions 7° x 2.4 (for ovals and rectangles) or 3.1° (for bowties), duration of 480 ms, Figure 1A,B), which was followed by a second variable delay period (500-900 ms). The width of the concrete cues was subjected to a small amount of jitter (+/- 0, 0.35°). The concrete cue was presented at either 340° ("north"), 60° ("southwest") or 125° ("southeast"). Finally, three identical white circle targets (1.72° diameter) were presented 12 degrees to the left, directly vertical, and right of the fixation cross. The monkey had to combine the information conveyed by the two rule cues to produce a saccade toward the correct target within 700 ms of targets onset. To obtain a reward, the gaze had to land in the correct target window (6° x 6°, centered on the target) and remain inside for 50 or 300 ms. When these conditions were met, correct feedback was signaled to the animal: juice reward delivered and an auditory tone played over a pair of speakers positioned on either side of the monitor. If the gaze left the fixation window, flew through the target window, or an incorrect target was foveated, then a distinct auditory tone sounded over the speakers and the trial was aborted. The duration of the intertrial interval was 1100-1700 ms.

Electrophysiological recording

We used LMAs with either 24 contacts, 12.5 µm contact diameter and 200 µm spacing between contacts (Microprobes for Life Science, Gaithersburg, MD), 32 contacts, 12.5 µm contact diameter and 150 µm spacing (Microprobes for Life Science; Microprobes LMAs MRI-compatible), or 32 contacts, 15 µm contact diameter and 150 µm spacing (Plexon V-Probes). LMAs were platinum/iridium and 0.8-1 MΩ impedance. We recorded electrical brain activity (filtered 0.1-7500 Hz and sampled at 40 kHz) using a preamplifier with a high impedance headstage and an OmniPlex data acquisition system controlled by

PlexControl software (Plexon). The system also recorded digitized eye position from the eye tracker and behavioral event information issued from our task control software.

Electrode array localization

The T1-weighted structural images with electrodes held *in situ* by the custom guide tubes allowed for confirmation of optimal linear electrode placement following craniotomy surgery (Figure 1D). Specifically, the electrode is not visible in the images, but a susceptibility "shadow" artifact appears along its length with a width of approximately one voxel (0.5 mm^3 on either side of the electrode). Probe tracks and target depths were calculated based on these images – the individual monkey's anatomy – with cross-referencing to stereotaxic atlases. We re-positioned electrodes as necessary and re-acquired T1-weighted structural volumes until electrodes were in their desired locations in the thalamus and cortex. Offline, we registered the images with electrodes *in situ* to the high-quality structural image acquired prior to surgery. These procedures enabled targeting of both banks of the PS in the aPFC and pPFC, the PS-connected MDc (putative MDpc) and SNr-innervated VAm (putative VAmc) (see Figure 1C,D).

We used the depth measurements from structural images to guide initial positions of the linear electrodes across the two PFC and thalamic recording sites. All contacts spanned approximately 4.6 mm. This enabled coverage across both banks of the PS (thus, acquisition of neural data from both area 46d and 46v) and a majority of the dorsal-to-ventral span of the thalamic nuclei. Online, we adjusted electrode position if necessary to maximize the number of contacts showing single- or multi-unit spiking activity. In the cortex, we used current source density (CSD) and analyses of unipolar local field potentials (LFPs) to assign the electrode contacts to cortical layers. We also performed post-mortem histology to reconstruct electrode tracks in one monkey, to corroborate evidence from MRI and electrophysiology. After fixing the brain in 10% neutral buffered formalin, the right hemisphere was sectioned into approximately 5 mm thick coronal slices, embedded in paraffin, then thinly sectioned (8 μm). In proximity to the ROIs, we stained sections with Hematoxylin and Eosin, and visualized sections under a microscope to confirm electrode tracks through our ROIs.

Neural data preprocessing

Spike data

We used two methods to isolate spiking activity from the acquired wideband data. (1) Using Plexon Offline Sorter software, we bandpass filtered data (250-5000 Hz) to extract potential spiking activity (Butterworth, order 4, zero-phase filter) for spike sorting. We used principal components analysis to extract features of the spike waveforms and a T-distribution expectation maximization algorithm to identify clusters of spikes with similar features. (2) Using Kilosort[50] running in Matlab, with manual curation steps taken in the Phy GUI (https://github.com/cortex-lab/phy/), we performed data preprocessing, spike clustering and template matching. We confirmed that the two methods isolated the same units by running each on the same recording session. We isolated activity from 1094 neurons (229 in aPFC, 340 in pPFC, 196 in VAm, 329 in MDc) across 51 recording sessions (19 in Monkey J, 32 in Monkey M).

Local field potentials

We applied a sixth-order Butterworth zero-phase filter to lowpass filter data to 100 Hz for current source density (CSD) analysis and to 200 Hz for relative power analysis. Next, data were downsampled to 1000 samples/second. We then applied linear detrending. To remove powerline noise, we used either the

Chronux function rmlinesc to remove significant sine waves from the LFP data for CSD, or an adaptive filter method[51]. The method began with the utilization of an adaptive notch filter to determine the primary frequency of the interference. Leveraging this identified frequency, we synthesized harmonics using discrete-time oscillators. Subsequently, we used a modified recursive least squares algorithm to estimate the amplitude and phase of each harmonic. We then subtracted and calculated interference from the original data to obtain a cleaner signal. Finally, LFP data were cut into chunks surrounding trial events of interest. To minimize the impact of motion artifacts, we excluded any trials with LFP amplitude greater than 5 times the standard deviation of the mean.

Current source density (CSD)

We designated electrode contacts and any isolated single units to either superficial (layers 1-3), middle (400-450 µm span above bottom of layer 4) or deep (layers 5 and 6) layers[14] based on inverse CSD analysis[52]. For this, we used the CSDplotter toolbox for Matlab for calculating the inverse CSD in response to concrete cues in the HRT (and this was confirmed with response to other task events). Linear electrodes measure the LFP, ϕ, at N different cortical depths/electrode contacts along the z axis with spacing h. The standard CSD, C_{st}, is estimated from the LFPs using the second spatial derivative.

$$C_{st}(z) = \frac{(\phi(z+h) - 2\phi(z) + \phi(z-h))}{h^2} \qquad \textbf{(Eq. 1)}$$

LFPs can also be estimated from CSDs (in matrix form $\Phi = F\hat{C}$ where Φ is the vector containing the N measurements of ϕ, $\hat{C}$ is the vector containing estimated CSDs, and F is an NxN matrix derived from the electrostatic forward calculation of LFPs from known current sources). The method uses the inverse of F to estimate the CSD, i.e., $\hat{C} = F^{-1}\Phi$. It is assumed that the CSD is stepwise constant between electrode contacts, so the sources are extended cylindrical boxes with radius R and height h.

$$F_{ji} = \int_{z_i - \frac{h}{2}}^{z_i + \frac{h}{2}} \frac{1}{2\sigma}\left(\sqrt{(z_j - z')^2 + R^2} - |z_j - z'|\right) dz' \qquad \textbf{(Eq. 2)}$$

Where σ is the electrical conductivity tensor and $\phi(z_j)$ is the potential measured at position z_j at the cylinder center due to a cylindrical current box with CSD, C_i, around the electrode position z_i. The inverse CSD provides an estimate of CSD around all N electrode contacts (contrast with standard CSD which does so for N-2) and does not require equidistant contacts like the standard CSD, which permits the analysis of sessions with noisy channels that should be removed.

Relative LFP power

We cross-referenced our layer designations estimated using CSD by calculating the relative LFP power across channels/contacts of our cortical LMAs. We analyzed LFP data during the 500 ms prior to abstract cue onset (last 500 ms of preparatory period), which should be weakly stationary. For each channel, we obtained the power spectrum (using the Chronux function mtspectrumc (with time bandwidth product (TW) = 2 and number of tapers (K) = 3 in Matlab), then normalized the power of that channel by that of the channel with the highest power for a given frequency. This revealed the expected common pattern for cortical relative power maps where activity at higher frequencies (50-150 Hz) had higher power in superficial channels than deep ones, while oscillatory activity in the alpha-beta frequencies (10-30 Hz)

had higher power in deep channels compared to superficial ones (Figure 1F), with the characteristic "swoosh" pattern previously described[15,16].

Wilcoxon rank-sum tests for selectivity analyses of spike data

We aimed to ascertain the proportion of selective cells in various regions including aPFC, pPFC, MD, and VA. To this end, we first calculated the spike counts for each defined interval of interest. Subsequently, we carried out a non-parametric Wilcoxon rank-sum test for every cell in each specific condition, namely, Abstract Rule, Concrete Rule, Saccade Direction, and Outcome.

To mitigate the challenges arising from multiple comparisons, we applied Holm's correction. A cell was classified as "selective" based on a final p-value less than 0.05, following the adjustment for multiple comparisons.

Furthermore, we determined the proportion of neurons exhibiting selectivity for the abstract rule condition over time, employing 100 ms running windows with a 50 ms overlap. This analysis invoked the same statistical test as previously described. To substantiate the significance of the calculated proportions, we executed a 10,000-step bootstrapping process, wherein we randomly chose 80% of the cells and assessed the proportion of selective cells in each iteration. Moreover, we contrasted the selectivity against a null distribution, derived from permutating the abstract rule labels and determining the selective cell proportions in each permutation.

To calculate the latency, we measured the times when the portion of selective cells reached 50% of their maximum selectivity across all 10,000 bootstrap samples. Then, we conducted a linear trend analysis using a General Linear Model (GLM) and linear contrasts.

Selectivity index analyses of spike data

Spike rate

We calculated spike density functions to measure the responses of the prefrontal cortical and thalamic neurons. For spike density functions and selectivity index analyses, spikes from each trial were convolved with a Gaussian pulse (λ = 20 ms) and we averaged the resulting functions to produce spike density functions for each neuron, with a temporal resolution of 4 ms.

Selectivity index formula

To characterize the response properties of neurons in thalamocortical circuits, we calculated selectivity indices (SIs[17,18])

$$SI = (n - \sum_{i=1}^{n}(\frac{\lambda i}{\lambda max}))/(n-1) \quad \textbf{(Eq. 3)}$$

where n is the number of conditions and λ is the spike rate in a given condition. We applied this index in a sliding window along 3.0 s chunks of spike data, aligned to various events of interest (window width = 100 ms; step size = 4 ms), thus obtaining a timecourse of SIs from the spike density functions for the abstract rule, concrete rule and direction. To define a selective neuron, we used a threshold which corresponded to baseline mean + 2 standard deviations above this baseline mean, and further the neuron had to exhibit sustained significant SI values for at least 12 consecutive bins (48 ms). We calculated *abstract rule and abstract cue SI* aligned to the abstract rule cue presentation onset, to assess cue-evoked sensory and rule selectivity as well as rule selectivity during the first delay period. We calculated *concrete*

rule and concrete cue SI aligned to the concrete rule cue presentation onset, to assess cue-evoked sensory and rule selectivity. We calculated *directional SI* aligned to the saccade onset to characterize preparatory, peri-saccade and post-saccade directional (saccade or target) selectivity. Baseline windows were the last 500 ms of the fixation period prior to abstract rule cue onset, for the abstract rule and delay 1, and the last 500 ms of delay 1 for events which followed delay 1, i.e., the concrete rule and target/saccade direction. To define latency to selectivity of a particular single neuron to a visual cue, we found the maximum SI value within the window of interest and calculated the time between event onset and time until the SI reached half of this maximum value. We excluded neurons with unrealistic latencies (i.e., <50 ms).

Defining sensory neurons for exclusion from abstract and concrete rule selectivity index analyses

Abstract cues

We included two cues for each abstract rule to enable the dissociation of visual neuronal responses from abstract rule neuronal responses. We ran a sensory SI where the spike rates associated with each of the four abstract cues were input. We reasoned that if a neuron was rule selective, the cues associated with its maximum two spike rates would reflect the same abstract rule (i.e., red circle and blue triangle, or green circle and yellow triangle). If a neuron was detected as selective based on the sensory SI and failed to meet this rule-related criterion, it was defined as a visual/sensory neuron and excluded from our subsequent abstract rule analyses.

Concrete cues

To define sensory responses to the concrete cues, we exploited the fact that (unlike the abstract rule cues), identical concrete rule cues (of which there are 9 permutations) are presented in both shape and orientation trials. We reasoned that if a cell has a maximal response to the same cue under both rule sets, that it is selective for that visual cue rather than any particular concrete rule. In this case, the neuron was excluded from our concrete rule analyses.

Population SI timecourse

We calculated the timecourse of SI and assessed for significant SI on individual neurons. If the neuron met criteria for a selective neuron in a particular analysis, we stored its SI data in a population matrix for further processing. For visualization of population SI, we calculated the average SI across each timebin (+/- standard error) and plotted its timecourse.

Statistical analyses of population SI timecourse

We compared the timing and magnitude of cognitive information represented by single neurons in the PFC and thalamus. To do this, we ran a univariate sliding window analysis (100 ms window width, stepped by 100 ms) of the population SI curves. Based on the thalamo-cortical model in Figure 1G that we envisioned supporting accurate performance of the HRT, we hypothesized that thalamic neuron populations would have stronger abstract rule selectivity than PFC populations (due to convergent inputs to the thalamus amplifying weak input signals). We used one-tailed t-tests to probe whether thalamic mean abstract rule SI values were greater than PFC mean SI values in 100 ms-wide nonoverlapping windows of the SI curves beginning 100 ms prior to abstract rule cue onset and continuing until the window beginning 800 ms after cue onset (corresponding to the shortest duration of delay 1, and common to all trials). For the concrete rule period, we used two-tailed t-tests, because we hypothesized that intracortical mechanisms and convergent cortico-thalamic mechanisms could give rise to different

magnitudes of concrete rule SI. We applied the same sliding window analysis for this period, as well. Finally, for the saccade aligned analysis of directional selectivity, we applied the same analysis to the three time windows: pre-saccade direction SI windows spanned from 700 ms until 100 ms prior to saccade onset; peri-saccade direction SI windows spanned from 100 ms prior to saccade until 100 ms following saccade onset; and post-saccade direction SI windows spanned from 100 ms until 900 ms post saccade onset.

Fano Factor Calculation

We employed the methodology delineated by Churchland and colleagues to compute the mean-matched Fano factor for each designated area[19]. This mean-matched approach controls for any spike rate differences between areas (as Fano factor decreases as spike rate increases). Our timeline of interest was centered around the abstract rule cue, spanning from 100ms before to 1000ms after the cue's onset. Within each area, we initially determined the Fano factor over a 100ms window, ensuring there was a 50ms overlap between successive windows. Following the mean-matching process, the survival rates of cells in these regions were observed to be notably high. Specifically, the aPFC region retained 84% of its cells, while the pPFC, MD, and VA regions had survival rates of 81%, 83%, and 82%, respectively. We found similar results with the mean-matched and raw (using all cells) Fano factor.

Pseudo-population neural decoding

To construct a robust and comprehensive training dataset and feature set for decoding purposes, we employed spike counts from each 5ms bin as input to our machine learning model. A pseudo-population of neurons was synthesized by: (1) incorporating neurons from all sessions into a consolidated vector, (2) for each neuron/condition combination (with labels abstract rule, concrete rule, saccade direction), a random trial from the session where the cell was located and the aforementioned condition was present was selected without replacement, and (3) this procedure was reiterated until we had chosen all trials for at least one session/neuron/condition. To effectively manage imbalances in the training data, the Synthetic Minority Over-sampling Technique (SMOTE) with a nearest neighbor approach was employed.

Our machine learning model is explicitly engineered to encapsulate both spatial (neuronal interactions) and temporal dependencies for the duration of a HRT trial, aligned to predetermined events (Abstract rule cue onset, Concrete rule cue onset, Saccade Onset). For extracting intricate spatial features from the spike counts of individual neurons and their combined patterns, we utilized a one-dimensional (1D) convolutional neural network (CNN) as the preliminary layer.

To ensure that the length of input to the succeeding layer remained constant, a max pooling layer was positioned after each CNN kernel. Our kernel dimensions were 3, 3, 5, and 7. Kernel sizes were meticulously chosen such that after passage through a 1x1 kernel, we had an outcome of 100 features as the input to our next machine learning module.

To account for the temporal dependencies of the firing pattern throughout the task, we employed a recurrent neural network (RNN) with a many-to-many architecture, which was designed to decode each condition using the enhanced features from the preceding layer. The proposed RNN comprises a Bidirectional Long-Short Term Memory (BiLSTM) layer, followed by a multiplicative local attention layer, a fully connected layer with dropout, and finally, a softmax layer for output. Given x(t) as the input at time t, the output of the BiLSTM forward path is computed as follows:

$$\mathrm{i}_{(t)} = \sigma\left(\mathrm{W}_{xi}^{T}.\mathrm{x}_{(t)} + \mathrm{W}_{hi}^{T}.\mathrm{h}_{(t-1)} + \mathrm{b}_{i}\right) \qquad (\mathbf{Eq.\,4})$$

$$\mathrm{f}_{(t)} = \sigma\left(\mathrm{W}_{xf}^{T}.\mathrm{x}_{(t)} + \mathrm{W}_{hf}^{T}.\vec{\mathrm{h}}_{(t-1)} + \mathrm{b}_{\mathrm{f}}\right) \qquad (\mathbf{Eq.\,5})$$

$$\mathrm{o}_{(t)} = \sigma\left(\mathrm{W}_{xo}^{T}.\mathrm{x}_{(t)} + \mathrm{W}_{ho}^{T}.\vec{\mathrm{h}}_{(t-1)} + \mathrm{b}_{\mathrm{o}}\right) \qquad (\mathbf{Eq.\,6})$$

$$\mathrm{g}_{(t)} = \tanh\left(\mathrm{W}_{xg}^{T}.\mathrm{x}_{(t)} + \mathrm{W}_{hg}^{T}.\vec{\mathrm{h}}_{(t-1)} + \mathrm{b}_{g}\right) \qquad (\mathbf{Eq.\,7})$$

$$\mathrm{c}_{(t)} = \mathrm{f}_{(t)} \otimes \mathrm{c}_{(t-1)} + \mathrm{i}_{(t)} \otimes \mathrm{g}_{(t)} \qquad (\mathbf{Eq.\,8})$$

$$\vec{\mathrm{h}}_{(\mathrm{t})} = \mathrm{o}_{(t)} \otimes \tanh\left(\mathrm{c}_{(t)}\right) \qquad (\mathbf{Eq.\,9})$$

Where $\mathrm{i}_{(t)}, \mathrm{f}_{(t)}, \mathrm{c}_{(t)}, \mathrm{o}_{(t)}$ are the input gate, forget gate, cell gate and output gate respectively. The weight matrices and bias vectors associated with the input, forget, cell, and output gates, as well as the hidden state at time *t* , are denoted by *W* (with appropriate subscripts) and *b* (with suitable subscripts). $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$, and the output of the backward path is $\overleftarrow{\mathrm{h}}_{(t)}$.

The attention layer output can be calculated as follows:

$$\mathrm{u}_{(t)} = \tanh\left(\mathrm{W}_{s}.[\vec{\mathrm{h}}_{(t)}; \overleftarrow{\mathrm{h}}_{(t)}] + \mathrm{b}_{s}\right) \qquad (\mathbf{Eq.\,10})$$

$$\alpha_{(t)} = \frac{\exp\left(\mathrm{u}_{(t)}^{T}\mathrm{u}_{w}\right)}{\sum_{t}\exp\left(\mathrm{u}_{(t)}^{T}\mathrm{u}_{w}\right)} \qquad (\mathbf{Eq.\,11})$$

$$\mathrm{s}_{(t)} = \sum_{t' \leq t} \alpha_{(t')} \left[\vec{\mathrm{h}}_{(t')}; \overleftarrow{\mathrm{h}}_{(t')}\right] \qquad (\mathbf{Eq.\,12})$$

Where:

- $\mathrm{u}_{(t)}$ is the activated attention weight at time *t*, computed through a tanh activation function applied to a linear combination of the forward and backward hidden states, $\vec{\mathrm{h}}_{(t)}$ and $\overleftarrow{\mathrm{h}}_{(t)}$, influenced by the attention weight matrix W_{s} and bias vector b_{s}.
- $\alpha_{(t)}$ represents the normalized attention weight at time *t*, derived by applying a softmax function over $\mathrm{u}_{(t)}$ to ensure that the sum of the attention weights over all time steps equals one. This is achieved by exponentiating $\mathrm{u}_{(t)}$ scaled by u_{w} and then dividing by the sum of the exponentiated $\mathrm{u}_{(t)}$ values over all time steps, giving a probability distribution over the time steps.
- $\mathrm{s}_{(t)}$ stands for the context vector at time *t*, calculated as a weighted sum of the concatenated forward and backward hidden states up to time *t*, using the normalized attention weights $\alpha_{(t')}$.
- W_{s} is the weight matrix associated with the attention mechanism, which learns to weigh the importance of different parts of the hidden states for the computation of the attention weights.
- b_{s} is the bias vector associated with the attention mechanism, complementing W_{s} in learning the attention weights.
- u_{w} is a learnable parameter that scales $\mathrm{u}_{(t)}$ before the softmax operation in the computation of $\alpha_{(t)}$ learning to adjust the scaling of $\mathrm{u}_{(t)}$ to control the sharpness of the distribution.

The output of the attention layer was fed to a fully connected layer with dropout followed by a softmax layer which results in class conditional probability as equation 13 specifies.

$$\hat{y}_{(t)} = \text{softmax}\left(\text{W}_{FC}.\text{s}_{(t)} + \text{b}_{FC}\right) \qquad (\textbf{Eq. 13})$$

The weights and biases for all BiLSTM, attention, and fully connected (FC) layers undergo adjustment through backpropagation in time, a key algorithmic process integral to training our neural network model. The hyperparameters of the model, which include the quantity of hidden units in the BiLSTM, Attention, and FC layers, the learning rate, dropout probability, and the learning rate of stochastic gradient descent among others, are refined through an exhaustive grid search cross-validation.

In evaluating the efficacy of our model, decoding accuracy was ascertained using a 10-fold cross-validation method. Here, 60% of trials were used as the training set and the remaining 40% as the test set. The stratification and shuffling techniques were employed during this division of data sets, ensuring that we minimized any potential accuracy bias that could arise due to imbalanced classes.

To ensure a thorough and unbiased selection of trials, pseudo-population data generation was conducted 10,000 times. Accuracy measurements for test data were performed for each cross-validation fold and for each pseudorandom instance of the dataset, providing us with a robust statistical overview of the model's performance.

The machine learning models were trained separately for each brain region of interest, including the aPFC, pPFC, MDc, and VAm. For the aPFC and pPFC in particular, we created four distinct models: one encompassing all cells and the remaining three models each representing superficial, middle, and deep layer cells.

Cortico-basal ganglia-thalamic model

To understand the underlying mechanism supporting the selection and maintenance of abstract rules, we built a layer-specific Cortico-Striatal-Thalamic model. Here, PFC includes three layers – superficial, middle, and deep. The trial starts with the abstract cue presentation (450 ms) followed by a delay (900 ms). As for the HRT, there were four cues that signal two abstract rules: shape or orientation. The blue triangle (BT) or red circle (RC) signals the shape rule, whereas the green circle (GC) or yellow triangle (YT) signals the orientation rule. This model simulates the flow of information through the PFC layers, striatum, substantia nigra pars reticulata and thalamus. Specific parameter, network connectivity and synaptic weight values are shown in Supplementary Tables 1 and 2.

Leaky integrate-and fire-model

We used the leaky integrate-and-fire model (LIF) to simulate the activity of individual neurons. LIF is a simplified spiking neuron model where the electrical properties of a neuron's membrane potential can be modeled with a parallel capacitor and a resistor. In this model, a neuron will fire when its membrane potential surpasses a certain threshold (-50 mv); after firing, its potential will reset (-65 mv) and cannot fire for a refractory time (1 ms). The time evolution of membrane potential (V) for neuron "j" is shown in Equation 14.

$$C\frac{dV^j}{dt} = \text{-g}(V^j - E_L) + I^j_{driv.} + I^j_{syn.} + I^j_{noise} \qquad (\textbf{Eq. 14})$$

In this equation, we are computing how the voltage across the neuron "j" membrane changes as a function of its intrinsic driving current ($I_{driv.}$), synaptic inputs from its presynaptic cells ($I_{syn.}$) and stochastic inputs

(I_{noise}). Here, "*C"* is the membrane capacitance (1 nF), "t" is time, "E_L" represents resting potential (-65 mv), and "g" is the membrane conductance (g = 1, for simplicity). The neuron time constant is 20 ms.

The synaptic current ($I_{syn.}$) to the neuron "J" could be described with the general form represented in Equation 15, as follows:

(Eq. 15)

$$I_{syn.}^{j} = \alpha W \sum_{j} A_{i,j}\big(H(\mathrm{t} - t_s - t_d) - H(\mathrm{t} - t_s - t_d - T)\big)$$

"W" stands for the synaptic weight of the incoming input. "$A_{i,j}$ " is the adjacency matrix that determines from which neurons "*i*" there are incoming connections to the neuron "*j*." Here, "H" is the *Heaviside* function controlling the synaptic current duration, "t_s" is the time of presynaptic spike from the neuron "i," "t_d" is the synaptic delay between the two neurons (5 ms in most cases). "T" is the duration of the pulse that the presynaptic cell sends to the postsynaptic cell (1.5 ms for excitatory driving currents and 100 ms for modulatory effect). Here, "α" represents the gain modulation factor from VA to PFC superficial cortical cells ("α" = 2). This factor is 1 for other connections. When the pre-synaptic neuron is inhibitory, this synaptic current has a negative sign, and the negative pulse drops exponentially to compensate for the burst-like activity of fast-spiking cells (Equation 16).

$$I_{syn.}^{j} = W \sum_{j} e^{-2\frac{t - t_s - t_d - 1}{T}} \big(H(t - t_s - t_d) - H(t - t_s - t_d - T)\big) \quad \textbf{(Eq. 16)}$$

Network structure

<u>PFC layers</u>

The first cells in the PFC that receive sensory signals and fire are the excitatory neurons in the middle layer (the middle layer consists of 50 neurons), which pass it to the superficial layer. Different populations of neurons in the middle layer get excited depending on the presented stimulus (e.g., BT vs. GC). For simplicity, we modeled starting cells in the superficial layer, which receives the sensory signal from the middle layer, in four chains (each chain has 50 neurons, a total of 200 superficial PFC cells) associated with each visual stimulus (BT, RC, GC and YT). Each neuron is connected locally to the neurons within and across the chains; connections within the chains are 1.3* stronger than across the chains. These superficial cells are also connected to a population of secondary PFC cells that get activated (and become rule-selective) later. These four superficial chains converge to four chains in the deep layer (similar to the superficial layer, each chain has 50 neurons, for 200 total deep layer PFC neurons) with a connectivity matrix that has the strongest weight on the diagonal (**2*, for example, from *superficial chain 1* [BT] to *deep chain 1*) and weaker weights off the diagonal (**0.2* from superficial chain 1 [BT] to deep chain 2, 3, 4). To investigate the effects of convergence from superficial to deep cortical layers, we also adjusted the weights off the diagonal that corresponded to the same rule (**0.2 and greater* from superficial chain 1 [BT] to deep chain 2 [RC]; *0.7 and *0.8 in Figure S10). This suggested that the range of plausible convergence from superficial to deep cortical layers does not generate rule selectivity in PFC early after abstract cue onset. Although stronger convergence might eventually lead to early, abstract rule selectivity in the cortex, it would not be consistent with the electrophysiological data (e.g., Figure 2L)

<u>Disinhibition of VA</u>

The PFC deep layer projects to the striatum, modeled with a population of inhibitory neurons. Elevated activity of the PFC deep layer causes firing in the striatum, inhibiting the inhibitory cells in the SNr. As SNr targets VA, this temporary decrease in SNr activity creates an excitable window for VA activity.

VA and MD thalamus

We modeled VA and MD thalamic nuclei. VA and MD cells integrate the synaptic inputs from their pre-synaptic deep layer PFC cells. The first two chains (1 and 2) from deep PFC converge onto VA and MD shape selective cells, and the second two chains (3 and 4) project onto MD and VA orientation selective cells. To be consistent with the known anatomical features of these circuits, in this model, connections projecting to the VA have stronger weights (*5) than those targeting MD. As mentioned above, the PFC-striatum-driven disinhibition of SNr allows for a facilitating time window for VA firing. The synchronous input from deep PFC recruits the VA and MD thalamic cells. VA gets activated first (t_d = 5 ms), as the time needed to reach MD is longer (t_d = 10 ms), due to the extra anatomical connection from cortical layer 5 to 6 (i.e., matrix cells, e.g., in VA, receive input from layer 5; and core cells, e.g., in MD, receive input from layer 6) and weaker modulatory PFC inputs. VA activation sends a broad pulse to their post-synaptic PFC cells, including the secondary PFC cells. VA cells have divergent modulatory (amplifying the gain of cortico-cortical connections by factor two) and excitatory effects on their post-synaptic PFC cells. The "secondary" population of PFC cells will receive driving and modulatory input from VA and driving current from MD. After activation, this population of PFC cells recruits more MD cells, which creates a PFC-MD loop to support rule information maintenance during the delay period. We added extended delay time within this PFC-MD loop to consider extra connections across layers that we are not modeling.

Cortical inhibition

In addition to VA's excitatory impact on superficial PFC cells, it modulates the connections within the PFC superficial layer and drives PFC inhibitory cells, resulting in a general inhibition. A more specific inhibition happens between the competing PFC *shape* and *orientation* ("secondary") neuronal ensembles. MD drives inhibitory cells, which target the wrong rule modality. This inhibition creates the "*winner-take-all*" paradigm that protects the fidelity of the signal.

Lesions of VAmc or MDpc

We model lesions of VA and MD by disconnecting them from the rest of the network.

Decoding Simulated Data

To decode abstract rules from simulated data in each of three conditions — namely, "healthy", MD lesion, and VA lesion — we employed the same machine learning model previously used for decoding of electrophysiological data. This model operates with a 50 ms time window for counting spikes. However, the method for generating the pseudopopulation of cells was adjusted.

For the simulated data, we randomly selected between 50% and 75% of cells at each step of randomization. Specifically, we chose a random integer from 50 to 75 to represent the percentage of cells to be selected. This selection was uniformly made from each area and under each abstract rule to ensure data balance. Decoding was then carried out using 10-fold cross-validation for each random iteration.

This process was repeated for 10,000 iterations across time, resulting in the generation of accuracy metrics.

Statistical analyses of decoding data

We scrutinized the decoding results of each ROI at each time bin. This was accomplished by comparing these results to a null distribution. This null distribution was derived through the permutation of class labels 10,000 times. Following this, we computed the accuracy at each instance to establish the null distribution.

To draw comparisons between the accuracies of different brain regions, a two-step statistical process was employed. Initially, the existence of an effect was ascertained using an analysis of variance (ANOVA). The ANOVA results, specifically F-statistics and p-values, were reported as effect size when necessary.

Post this, for each hypothesis a detailed contrast analysis was designed and conducted for each time bin. Below is an example of contrast design for one of the hypotheses: To rigorously test the specified hypotheses regarding the conveyance of abstract rule information among the various ROIs, namely VAm, MDc, pPFC and aPFC, the following series of contrasts was designed and implemented.

For the first hypothesis positing that the VAm population carries more abstract rule information than other areas, the contrast was structured as follows: [VAm - (MDc + pPFC + aPFC)/3]. This contrast is representative of the differential encoding of abstract rule information between VAm and the mean of MDc, pPFC and aPFC. A positive value from this contrast would support the hypothesis, suggesting that the VAm carries more abstract rule information than the other regions.

The second hypothesis asserts that the MDc population carries more abstract rule information than both the pPFC and aPFC. This was investigated through the following contrast: [MDc - (pPFC + aPFC)/2]. This contrast represents the difference in abstract rule information encoding between MDc and the mean of the pPFC and aPFC. A positive value here would support the hypothesis, suggesting that the MDc carries more abstract rule information than the pPFC and aPFC.

Lastly, for the third hypothesis stating that the pPFC carries more abstract rule information than the aPFC, we utilized the contrast [pPFC - aPFC]. A positive value from this contrast would support the hypothesis, indicating that the pPFC encodes more abstract rule information than the aPFC.

In more succinct notation, the designed contrasts for hypotheses tested and reported in the main text for abstract rule decoding are:

1. For the hypothesis that the VAm population carries abstract rule information earliest, the contrast is set as: [0, 0, 0, 1] for [aPFC, pPFC, MDc, VAm] respectively.
2. For the hypothesis that thalamic populations VAm and MDc carry rule information earlier than pPFC, the contrast is set as: [0, -1, 1, 1] for [aPFC, pPFC, MDc, VAm] respectively.
3. For the hypothesis that the VAm population carries more abstract rule information than the other ROIs 0-450 ms after abstract rule cue onset, the contrast was set as: [-1, -1, -1, 3] for [aPFC, pPFC, MDc, VAm] respectively.
4. For the hypothesis that MDc carries more abstract rule information than both pPFC and aPFC 0-450 ms after abstract rule cue onset, the contrast was designed as: [-1, -1, 2, 0] for [aPFC, pPFC, MDc, VAm] respectively.

5. For the hypothesis that pPFC carries more abstract rule information than aPFC 0-450 ms after abstract rule cue onset, the contrast was set as: [-1, 1, 0, 0] for [aPFC, pPFC, MDc, VAm] respectively.

Each of these contrasts, carried out for each time bin, were corrected for multiple comparisons using the Bonferroni method. These results were then reported with the respective t-statistics and p-values where necessary, highlighting the evidence supporting these hypotheses.

Finally, the latency for each ROI was calculated. This was accomplished by utilizing the decoding accuracy results and determining the time at which each region possessed decodable information, i.e., significantly above chance levels, as compared to the null accuracy distribution.

ACKNOWLEDGMENTS

This work supported by NIH grants R01MH110311, R01NS117901 and P51OD011106; Whitehall Foundation grant 2015-12-71; and Brain and Behavior Research Foundation grant 23017.

AUTHOR CONTRIBUTIONS

Conceptualization: J.P., Y.S.; Investigation: J.P., N.K., M.R., S.Moh., A.C., Y.S.; Methodology, Software, Formal Analysis: J.P., M.A., S.S., E.R., X.C., M.F., S.Mof., Y.S.; Writing – Original Draft: J.P., M.A., S.Mof. and Y.S.; Writing – Review and Editing: J.P., M.A., N.K., M.R., S.S., E.R., X.C., M.F., S.Moh., A.C., C.M., S.Mof., Y.S.; Supervision, Funding Acquisition: Y.S..

COMPETING INTEREST DECLARATION

The authors declare no competing interests.

REFERENCES

1 Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585-590 (2013). https://doi.org:10.1038/nature12160

2 Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci* **2**, 820-829 (2001). https://doi.org:10.1038/35097575

3 Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol* **37**, 66-74 (2016). https://doi.org:10.1016/j.conb.2016.01.010

4 Phillips, J. M., Kambi, N. A., Redinbaugh, M. J., Mohanta, S. & Saalmann, Y. B. Disentangling the influences of multiple thalamic nuclei on prefrontal cortex and cognitive control. *Neurosci Biobehav Rev* **128**, 487-510 (2021). https://doi.org:10.1016/j.neubiorev.2021.06.042

5 Jones, E. G. *The thalamus*. Second edition. edn, (Cambridge University Press, 2007).

6 Bolkan, S. S. *et al.* Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat Neurosci* **20**, 987-996 (2017). https://doi.org:10.1038/nn.4568

7 Schmitt, L. I. *et al.* Thalamic amplification of cortical connectivity sustains attentional control. *Nature* **545**, 219-223 (2017). https://doi.org:10.1038/nature22073

8 Garcia-Cabezas, M. A., Martinez-Sanchez, P., Sanchez-Gonzalez, M. A., Garzon, M. & Cavada, C. Dopamine innervation in the thalamus: monkey versus rat. *Cereb Cortex* **19**, 424-434 (2009). https://doi.org:10.1093/cercor/bhn093

9 Ilinsky, I. A. & Kultas-Ilinsky, K. Fine structure of the magnocellular subdivision of the ventral anterior thalamic nucleus (VAmc) of Macaca mulatta: I. Cell types and synaptology. *J Comp Neurol* **294**, 455-478 (1990). https://doi.org:10.1002/cne.902940313

10 Ilinsky, I. A., Jouandet, M. L. & Goldman-Rakic, P. S. Organization of the nigrothalamocortical system in the rhesus monkey. *J Comp Neurol* **236**, 315-330 (1985). https://doi.org:10.1002/cne.902360304

11 Kultas-Ilinsky, K. & Ilinsky, I. A. Fine structure of the magnocellular subdivision of the ventral anterior thalamic nucleus (VAmc) of Macaca mulatta: II. Organization of nigrothalamic afferents as revealed with EM autoradiography. *J Comp Neurol* **294**, 479-489 (1990). https://doi.org:10.1002/cne.902940314

12 Badre, D., Hoffman, J., Cooney, J. W. & D'Esposito, M. Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nat Neurosci* **12**, 515-522 (2009). https://doi.org:10.1038/nn.2277

13 Badre, D. & D'Esposito, M. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *J Cogn Neurosci* **19**, 2082-2099 (2007). https://doi.org:10.1162/jocn.2007.19.12.2082

14 Redinbaugh, M. J. *et al.* Thalamus Modulates Consciousness via Layer-Specific Control of Cortex. *Neuron* **106**, 66-75 e12 (2020). https://doi.org:10.1016/j.neuron.2020.01.005

15 Bastos, A. M., Loonis, R., Kornblith, S., Lundqvist, M. & Miller, E. K. Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *Proc Natl Acad Sci U S A* **115**, 1117-1122 (2018). https://doi.org:10.1073/pnas.1710323115

16 Mendoza-Halliday, D. *et al.* A ubiquitous spectrolaminar motif of local field potential power across the primate cortex. *Nat Neurosci* **27**, 547-560 (2024).

17 Wirth, S. *et al.* Single neurons in the monkey hippocampus and learning of new associations. *Science* **300**, 1578-1581 (2003). https://doi.org:10.1126/science.1084324

18 Moody, S. L., Wise, S. P., di Pellegrino, G. & Zipser, D. A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. *J Neurosci* **18**, 399-410 (1998). https://doi.org:10.1523/JNEUROSCI.18-01-00399.1998

19 Churchland, M. M. *et al.* Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat Neurosci* **13**, 369-378 (2010). https://doi.org:10.1038/nn.2501

20 Watanabe, Y. & Funahashi, S. Thalamic mediodorsal nucleus and working memory. *Neurosci Biobehav Rev* **36**, 134-142 (2012). https://doi.org:10.1016/j.neubiorev.2011.05.003

21 Phillips, J. M. *et al.* Topographic organization of connections between prefrontal cortex and mediodorsal thalamus: Evidence for a general principle of indirect thalamic pathways between directly connected cortical areas. *Neuroimage* **189**, 832-846 (2019). https://doi.org:10.1016/j.neuroimage.2019.01.078

22 Yasuda, M., Yamamoto, S. & Hikosaka, O. Robust representation of stable object values in the oculomotor Basal Ganglia. *J Neurosci* **32**, 16917-16932 (2012). https://doi.org:10.1523/JNEUROSCI.3438-12.2012

23 DeNicola, A. L., Park, M. Y., Crowe, D. A., MacDonald, A. W., 3rd & Chafee, M. V. Differential Roles of Mediodorsal Nucleus of the Thalamus and Prefrontal Cortex in Decision-Making and State Representation in a Cognitive Control Task Measuring Deficits in Schizophrenia. *J Neurosci* **40**, 1650-1667 (2020). https://doi.org:10.1523/JNEUROSCI.1703-19.2020

24 Sanchez-Gonzalez, M. A., Garcia-Cabezas, M. A., Rico, B. & Cavada, C. The primate thalamus is a key target for brain dopamine. *J Neurosci* **25**, 6076-6083 (2005). https://doi.org:10.1523/JNEUROSCI.0968-05.2005
25 Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci* **7**, 404-410 (2004). https://doi.org:10.1038/nn1209
26 Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181-1185 (2003). https://doi.org:10.1126/science.1088545
27 Badre, D. & D'Esposito, M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci* **10**, 659-669 (2009). https://doi.org:10.1038/nrn2667
28 Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S. & D'Esposito, M. Causal evidence for frontal cortex organization for perceptual decision making. *Proc Natl Acad Sci U S A* **113**, 6059-6064 (2016). https://doi.org:10.1073/pnas.1522551113
29 Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661-1663 (2014). https://doi.org:10.1038/nn.3862
30 Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782-790 (2012). https://doi.org:10.1016/j.neuroimage.2011.09.015
31 Steiskal, E. O. & Tanner, J. E. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics* **42** (1965).
32 Reeder, S. B., Hargreaves, B. A., Yu, H. & Brittain, J. H. Homodyne reconstruction and IDEAL water-fat decomposition. *Magn Reson Med* **54**, 586-593 (2005). https://doi.org:10.1002/mrm.20586
33 Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23 Suppl 1**, S208-219 (2004). https://doi.org:10.1016/j.neuroimage.2004.07.051
34 Woolrich, M. W. *et al.* Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45**, S173-186 (2009). https://doi.org:10.1016/j.neuroimage.2008.10.055
35 Ilinsky, I. A. & Kultas-Ilinsky, K. Sagittal cytoarchitectonic maps of the Macaca mulatta thalamus with a revised nomenclature of the motor-related nuclei validated by observations on their connectivity. *J Comp Neurol* **262**, 331-364 (1987). https://doi.org:10.1002/cne.902620303
36 Erickson, S. L. & Lewis, D. A. Cortical connections of the lateral mediodorsal thalamus in cynomolgus monkeys. *J Comp Neurol* **473**, 107-127 (2004). https://doi.org:10.1002/cne.20084
37 Goldman-Rakic, P. S. & Porrino, L. J. The primate mediodorsal (MD) nucleus and its projection to the frontal lobe. *J Comp Neurol* **242**, 535-560 (1985). https://doi.org:10.1002/cne.902420406
38 Xiao, D. & Barbas, H. CIrcuits through prefrontal cortex, basal ganglia, and ventral anterior nucleus map pathways beyond motor control. *Thalamus & Related Systems* **2**, 325-343 (2004).
39 Jezzard, P. & Balaban, R. S. Correction for geometric distortion in echo planar images from B0 field variations. *Magn Reson Med* **34**, 65-73 (1995). https://doi.org:10.1002/mrm.1910340111
40 Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med Image Anal* **5**, 143-156 (2001). https://doi.org:10.1016/s1361-8415(01)00036-6
41 Andersson, J. L. R. & Sotiropoulos, S. N. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* **125**, 1063-1078 (2016). https://doi.org:10.1016/j.neuroimage.2015.10.019
42 Paxinos, G., Huang, X. F. & Toga, A. W. *The rhesus monkey brain in stereotaxic coordinates*. (Academic Press, 2000).
43 Saleem, K. S. & Logothetis, N. K. *A combined MRI and histology atlas of the rhesus monkey brain in stereotaxic coordinates*. (Academic, 2007).
44 Olszewski, J. *The thalamus of the Macaca mulatta : an atlas for use with the stereotaxic instrument*. (S. Karger, 1952).

45 Behrens, T. E., Berg, H. J., Jbabdi, S., Rushworth, M. F. & Woolrich, M. W. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *Neuroimage* **34**, 144-155 (2007). https://doi.org:10.1016/j.neuroimage.2006.09.018

46 Behrens, T. E. *et al.* Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci* **6**, 750-757 (2003). https://doi.org:10.1038/nn1075

47 Klein, J. C. *et al.* Topography of connections between human prefrontal cortex and mediodorsal thalamus studied with diffusion tractography. *Neuroimage* **51**, 555-564 (2010). https://doi.org:10.1016/j.neuroimage.2010.02.062

48 Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X. & Kastner, S. The pulvinar regulates information transmission between cortical areas based on attention demands. *Science* **337**, 753-756 (2012). https://doi.org:10.1126/science.1223082

49 Behrens, T. E. *et al.* Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn Reson Med* **50**, 1077-1088 (2003). https://doi.org:10.1002/mrm.10609

50 Pachitariu, M., Steinmetz, N. A., Kadir, S. N., Carandini, M. & Harris, K. D. Fast and accurate spike sorting of high-channel count probes with KiloSort. *Advances in Neural Information Processing Systems* **29**, 4448-4456 (2016).

51 Keshtkaran, M. R. & Yang, Z. A fast, robust algorithm for power line interference cancellation in neural recording. *J Neural Eng* **11**, 026017 (2014). https://doi.org:10.1088/1741-2560/11/2/026017

52 Pettersen, K. H., Devor, A., Ulbert, I., Dale, A. M. & Einevoll, G. T. Current-source density estimation based on inversion of electrostatic forward solution: effects of finite extent of neuronal activity and conductivity discontinuities. *J Neurosci Methods* **154**, 116-133 (2006). https://doi.org:10.1016/j.jneumeth.2005.12.005

FIGURES

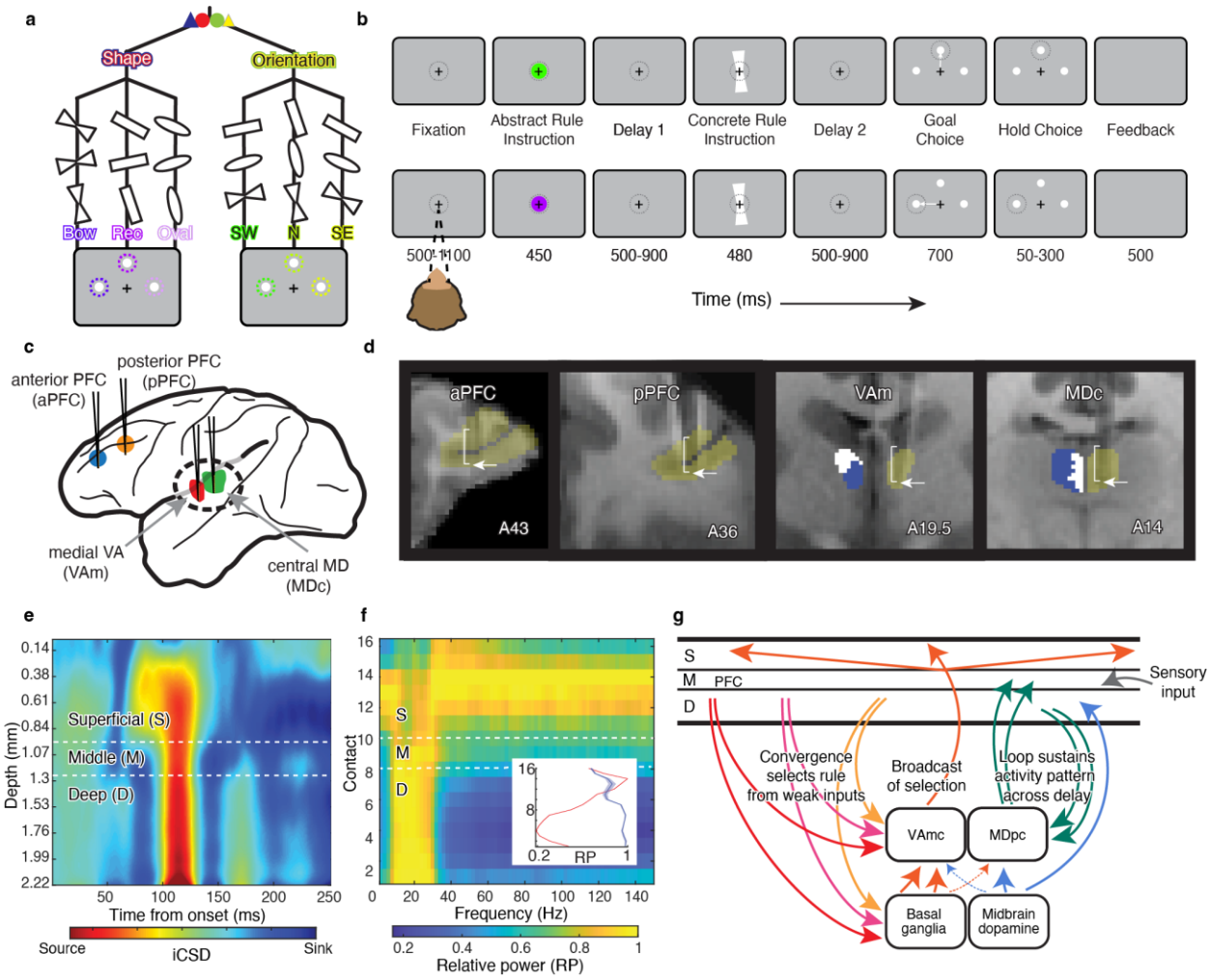

Figure 1. *Experimental design and hypotheses*. **a.** HRT stimulus permutations. **b.** HRT trial structure. Two trials with identical concrete rule stimuli but different preceding abstract rule instructions require two distinct stimulus-response mappings to obtain reward. Although we used blue and red cues to instruct shape, and green and yellow cues to instruct orientation, in **b** (and elsewhere) we represent shape instruction using purple and orientation instruction using yellowish green. In top row, abstract cue instructs that shape is relevant, and because the concrete rule following delay 1 is a bowtie, monkeys select leftward target. In bottom row, abstract cue instructs that orientation is relevant, and because this same concrete cue is oriented "north", monkeys select vertical target. **c.** Lateral view of monkey brain showing location of ROIs. **d.** Coronal structural MRIs showing probes in situ. Blue voxels show SNr projection zones in VAm and PFC projection zones in MDc (only displayed for left hemisphere, to allow probe visualization in right hemisphere), obtained using probabilistic tractography. Bracket indicates span of LMA contacts (~4.6 mm) and arrow indicates probe tip. **e,f.** Current source density (CSD) and relative LFP power enabled separation of LMA contacts into superficial, middle and deep layers. **g.** Projections to and from VAmc and MDpc with PFC layers. Ratio of matrix to core neurons is 50/50 for VAmc and 20/80 for MDpc.

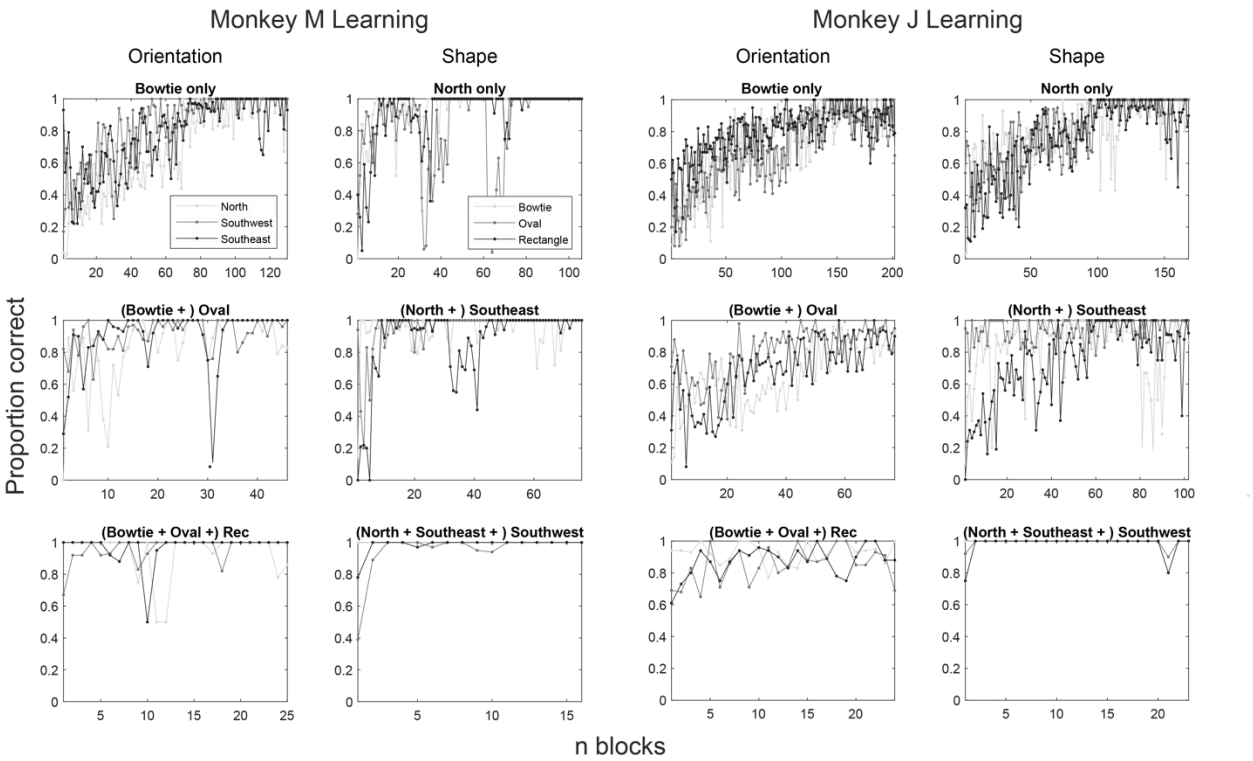

Figure S1. *Learning curves for the two monkeys during training on the HRT*. Under each abstract rule, monkeys began training with concrete cues associated with only one irrelevant feature that varied only in the relevant dimension (e.g., bowtie as the single irrelevant shape feature while training on the orientation rule set with all three orientations). As is typical for non-human primate learning of cognitive tasks, learning occurred slowly through operant mechanisms (see top row). However, when concrete cues associated with a second irrelevant feature were added, learning occurred much more quickly (see second row). When the cues associated with the final irrelevant feature were added, generalization of the rule set occurred immediately (see third row). Datapoints represent accuracy measured over the course of 100-trial blocks.

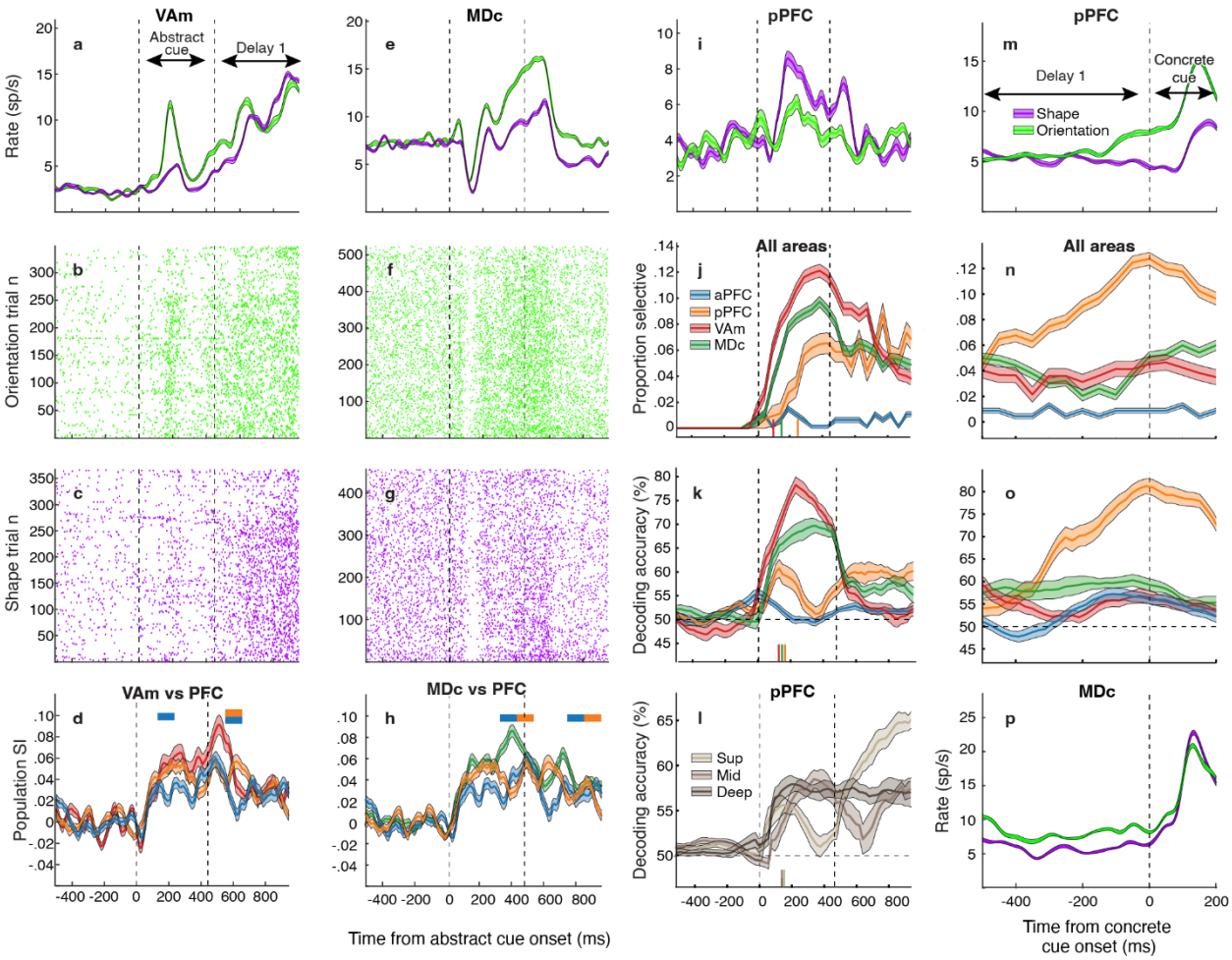

Figure 2. *Abstract-rule selectivity first in thalamus*. **a-l**, data and results aligned to onset of rule cue and beginning of delay 1; **m-p**, data and results aligned to onset of concrete rule cue (end of delay 1). **a-h**, Spike density functions (SDFs), raster plots and population SI timecourses for (a-d) VAm and (e-h) MDc neurons. Colored horizontal bars indicate significantly different 100-ms windows for each PFC region ($p < 0.05$, one-tailed t-tests). **i.** pPFC SDF. **j.** Proportion of rule-selective neurons in each ROI calculated in sliding 100-ms window. **k,l.** Timecourses of decoding accuracy for each (**k**) ROI or (**l**) pPFC layer. **m.** pPFC SDF. **n.** Proportion of rule-selective neurons in each ROI calculated in sliding 100-ms window. **o.** Timecourse of decoding accuracy for each ROI**. p.** MDc SDF.

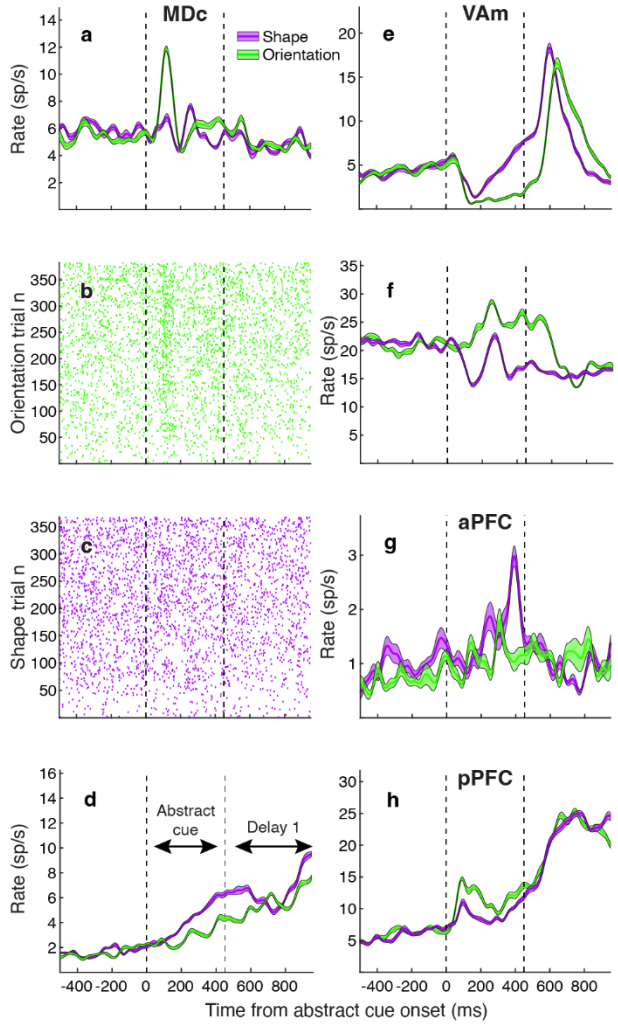

Figure S2. *Comparison of abstract-rule selective neurons in thalamus and cortex*. **a-c.** MDc SDF (a) and corresponding raster plots for orientation (b) and shape (c) showing abstract rule selectivity at onset. **d.** MDc neuron showing sustained selectivity. **e,f.** Two VAm neurons showing selectivity starts early. **g,h.** Selective cortical neurons in aPFC (g) and pPFC (h). Although the aPFC neuron in (g) shows rule-selectivity, abstract-rule selective neurons were less common in aPFC compared with other areas (Fig. 2j).

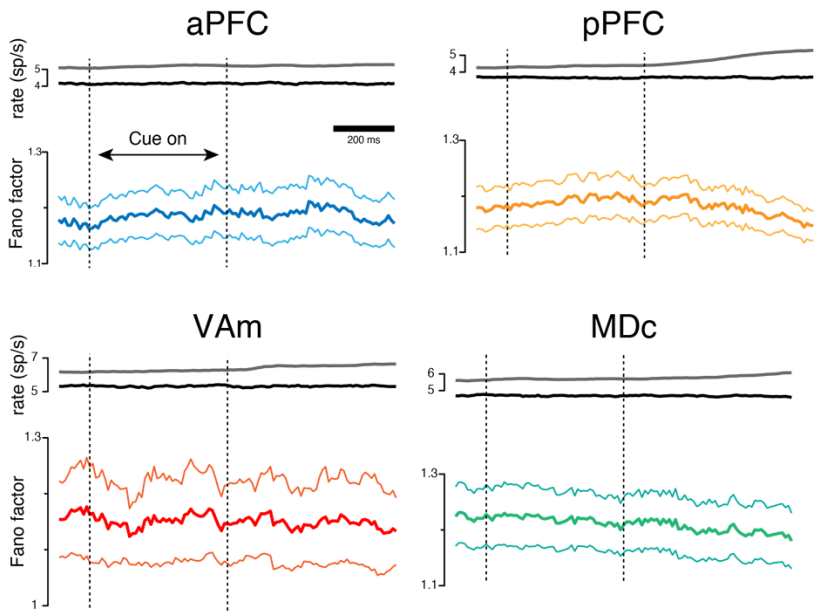

Figure S3. *Fano Factor does not differ between ROIs*. Population mean firing rate after the mean-matching process in black, at top of each ROI's panel; raw mean firing rate, prior to mean-matching, in grey. Mean-matched Fano factor calculated in sliding 100-ms windows, with 50-ms overlap between successive windows, at bottom of each ROI's panel. Firing rate and Fano factor from 100 ms before to 1000 ms after abstract rule cue onset.

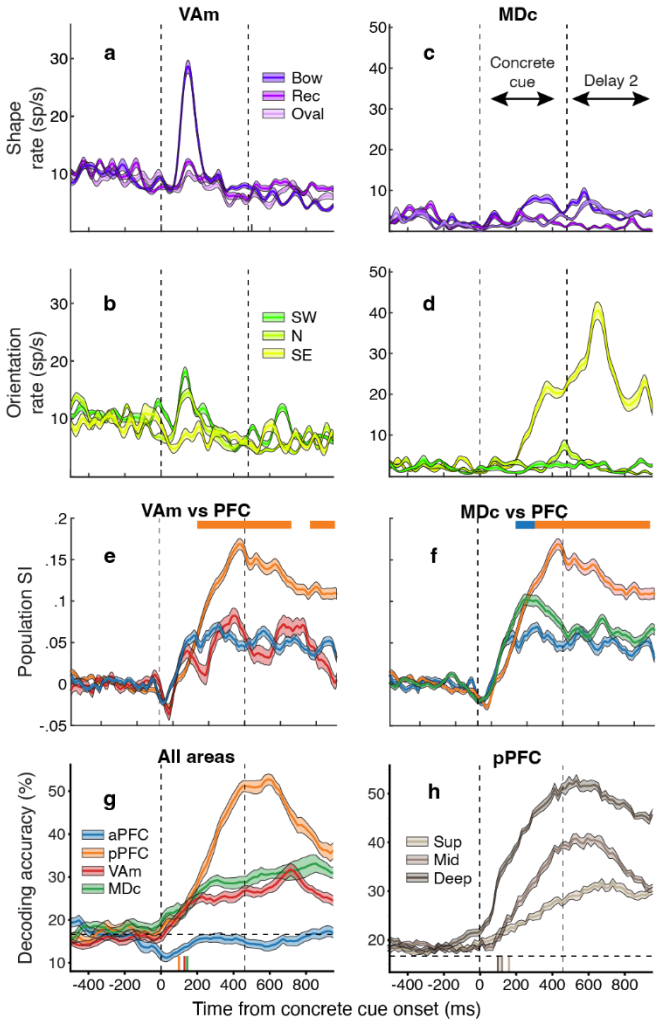

Figure 3. *Concrete-rule selectivity arises first in pPFC*. **a-d.** SDFs for single VAm (a,b) and MDc (c,d) neurons. **e,f.** Population SI timecourses for VAm (e) and MDc (f), compared with those from PFC. Colored horizontal bars indicate significantly different 100-ms windows for aPFC ($p < 0.02$) and pPFC ($p < 0.03$, 2-tailed t-tests). **g,h.** Timecourses of decoding accuracy for each ROI (g) or pPFC layer (h).

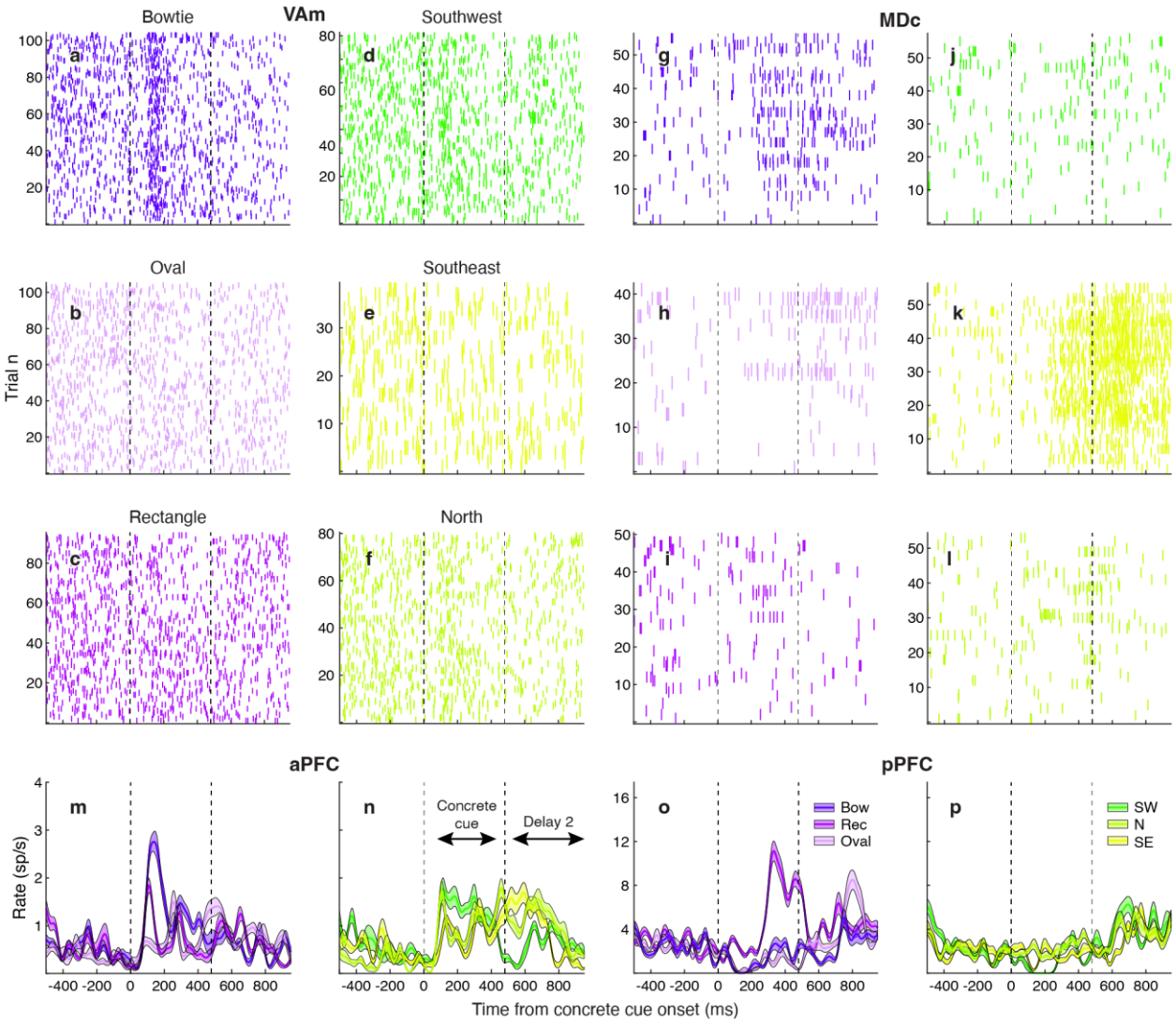

Figure S4. *Comparison of concrete-rule selective neurons in thalamus and cortex*. **a-l.** Raster plots associated with single neuron examples in Figure 3a,b (a-f) and Figure 3c,d (g-l). **m-p.** SDFs of neurons in aPFC (m,n) and pPFC (o,p).

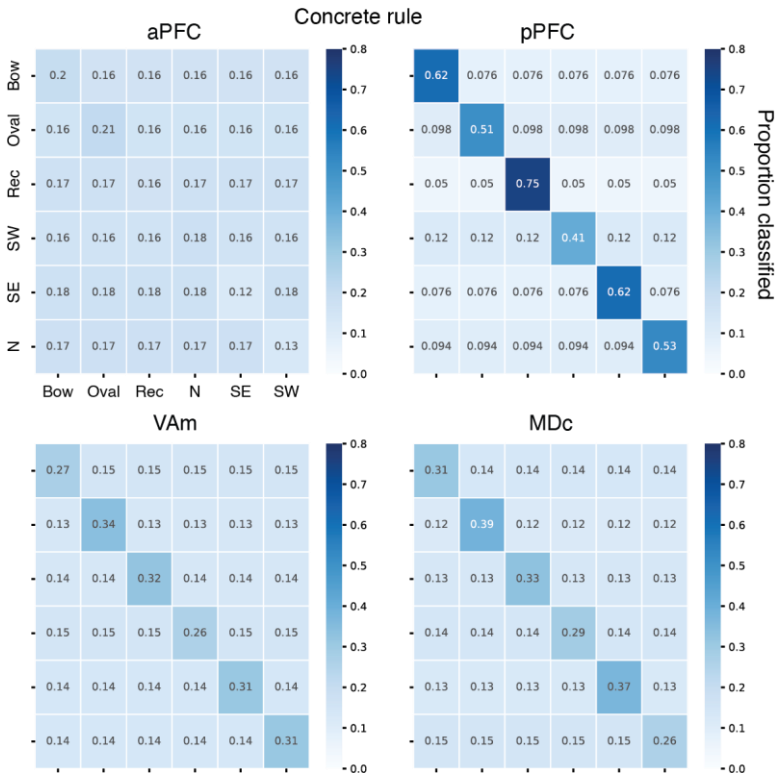

Figure S5. *Confusion matrices for concrete rule decoding.* Decoding accuracy color-coded. Darker diagonal indicates decoding of each of the six concrete rules.

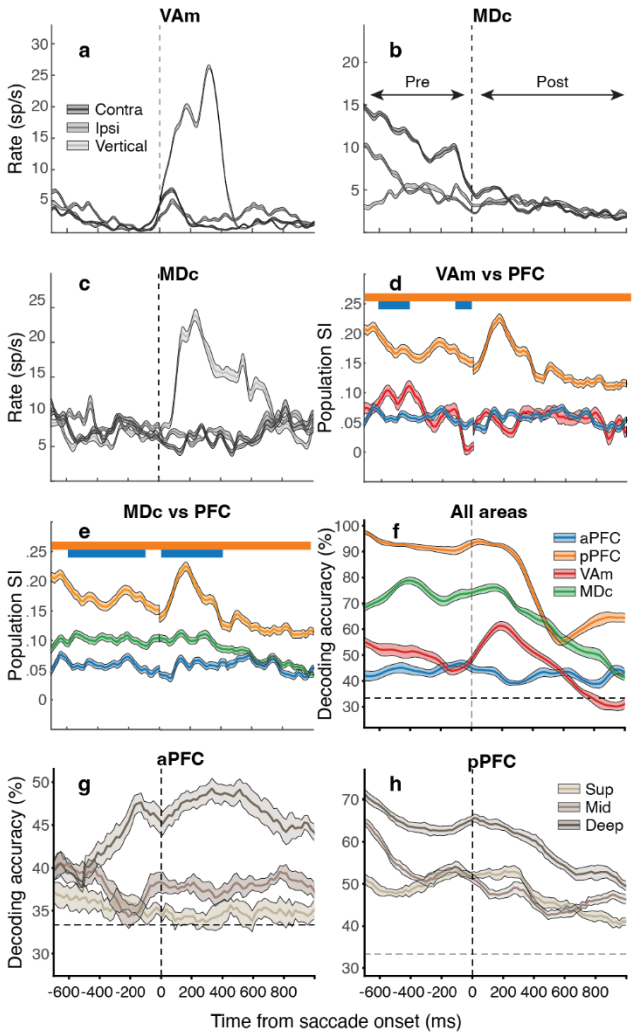

Figure 4. *Directional selectivity strongest in pPFC then MD*. **a-c.** SDFs for VAm (a) and MDc (b,c) neurons. **d,e.** Population SI timecourses for VAm (d) and MDc (e), compared with those from PFC. Colored horizontal bars indicate significantly different 100-ms windows for aPFC ($p < 0.05$) and pPFC ($p < 0.01$, two-tailed t-tests). **f-h.** Timecourses of decoding accuracy for each ROI (f) or PFC layer (g,h).

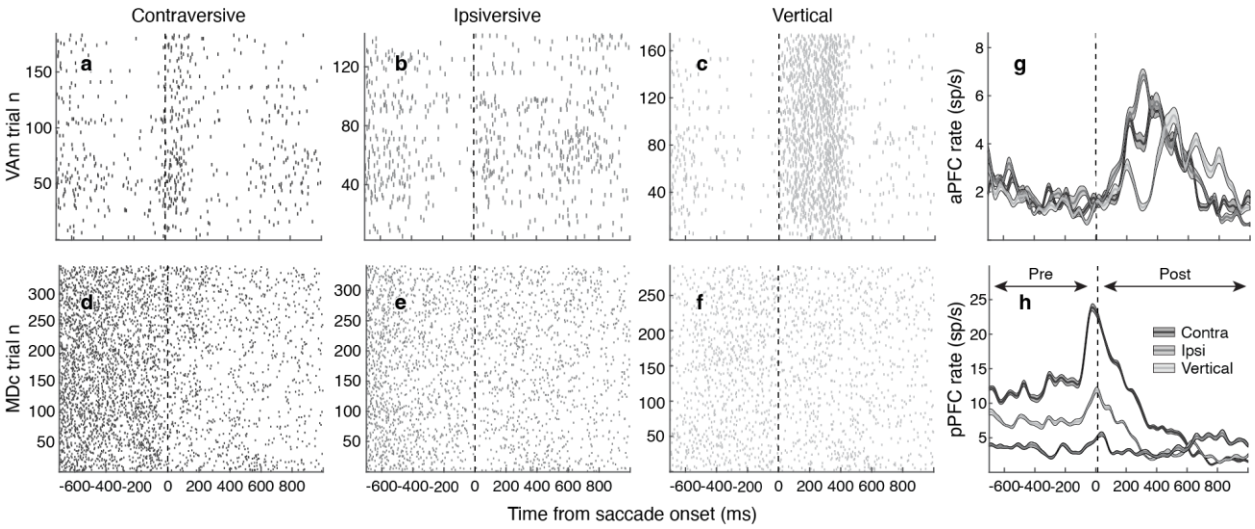

Figure S6. *Comparison of direction-selective neurons in thalamus and cortex*. **a-f.** Raster plots associated with single neuron examples in Figure 4a (a-c) and Figure 4b (d-f). **g,h.** SDFs of aPFC (g) and pPFC (h) neurons.

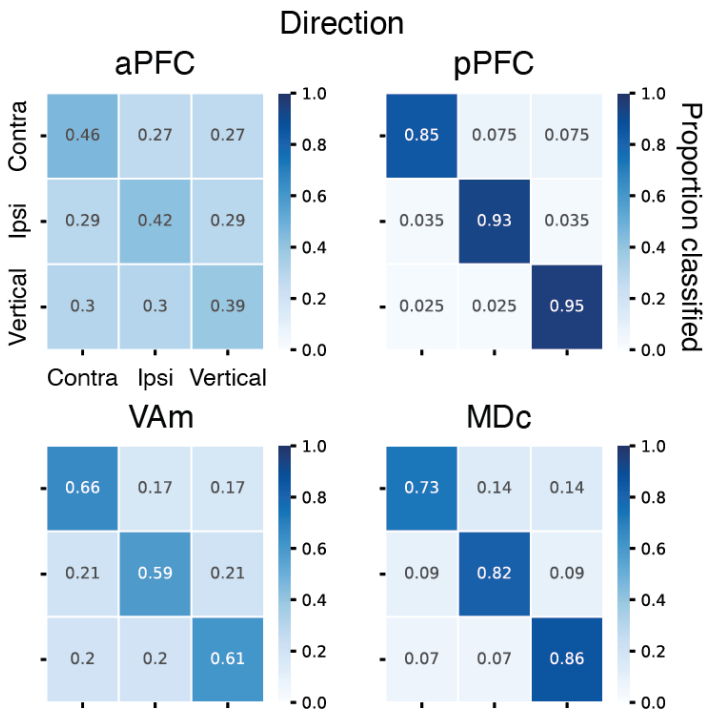

Figure S7. *Confusion matrices for direction decoding*. Darker diagonal indicates decoding of each of the three directions.

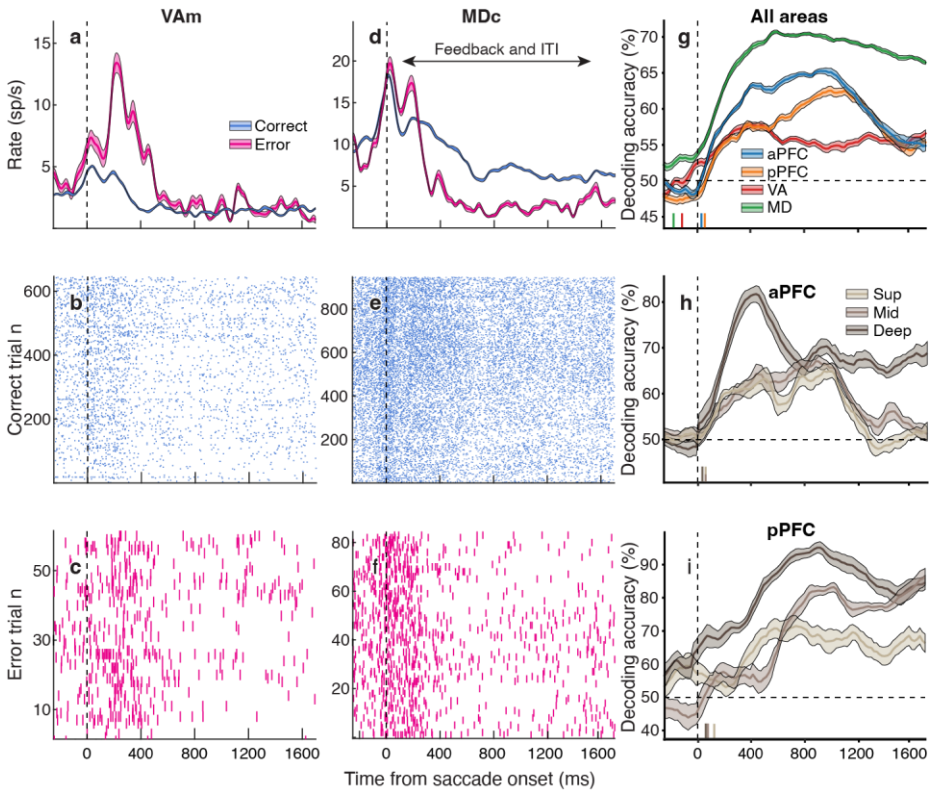

Figure 5. *Outcome selectivity arises first in thalamus*. **a-f.** SDF and raster plots for single VAm (a-c) and MDc (d-f) neurons. **g-i.** Timecourses of decoding accuracy for each ROI (g) or PFC layer (h,i).

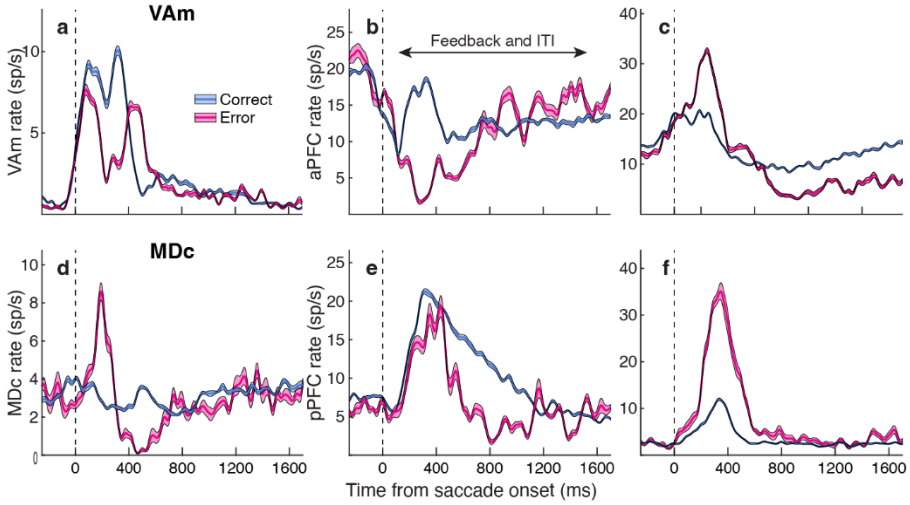

Figure S8. *Selectivity for trial outcome in thalamus*. **a-f.** SDFs of VAm (a-c) and MDc (d-f) neurons.

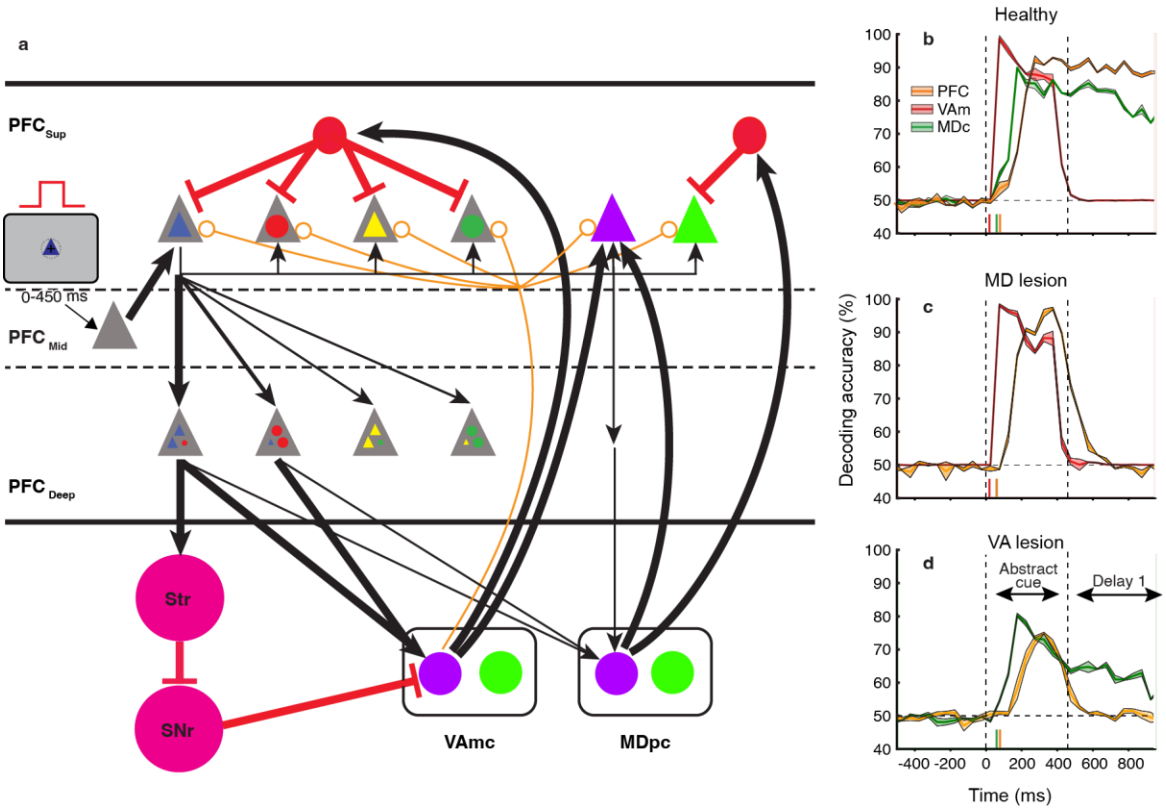

Figure 6. *PFC-BG-thalamic model reproduced* in vivo *abstract rule selectivity and thalamic-lesion modeling disrupted PFC rule representations*. **a.** *in silico* PFC includes superficial, middle (rule cue input layer) and deep layers; BG includes striatum and SNpr. VAmc receives disinhibitory input from BG and driving input from PFC deep layers; MDpc receives weaker, reflecting modulatory, input from PFC deep layers. VAmc provides divergent (matrix) and driving (core) input to PFC, and MDpc provides driving input to PFC, reflecting known primate anatomy and physiology[4]. Flow of information through PFC layers, striatum, SNr and thalamus simulated using leaky integrate-and-fire cells. Each triangle or circle represents a different cell population. PFC superficial layer contains populations selective for one of the four cues (blue triangle, red circle, yellow triangle, dark green circle) or one of the abstract rules (purple, shape rule; green, orientation rule). A population of inhibitory interneurons, driven by VAmc, imposed general inhibition on cue-selective PFC superficial cells; and a population, driven by MDpc, specifically inhibited superficial cells selective for the irrelevant rule. PFC deep layer populations receive driving input from superficial layer cells. VAmc and MDpc cell populations selective for shape (purple) or orientation (green) rule. Black arrows represent excitatory projections, red lines with flat terminations indicate inhibitory projections, and orange lines with open circle terminations represent divergent matrix projections. Thicker lines represent projections having synaptic weight > 0.003; thinner lines represent projections having synaptic weight ≤ 0.003 (model synaptic weights provided in Table S2). **b-d.** Time courses of abstract rule decoding accuracy from each ROI's spike data generated by complete, "healthy" (b), MD-lesioned (c) or VA-lesioned (d) models.

Supplementary Table 1. Model parameters and connectivity patterns

Description	**Value**
Neuron time constant	20 ms
Step time size	0.01 ms
Synaptic delay	5 ms
MD synaptic delay	10 ms
Firing threshold	-50 mv
Neuron resting potential	-65 mv
Membrane resistance	10 $\Omega.cm^2$
Membrane conductance	1 nF
Refractory time	1 ms
Pulse width for most PFC cells	1.5 ms
Pulse width for general inhibition	100 ms
Pulse width for winner-take-all	1 s
Window width for VA	100 ms
Connectivity Patterns	
Middle layer PFC cells to superficial	All-to-all connectivity in sequential blocks of ten; the last ten neurons are connected to the first ten neurons.
Within superficial PFC cells	All-to-all connectivity in sequential blocks of ten, within and across chains; the last ten neurons are connected to the first ten neurons.
Within deep layer PFC cells	All-to-all connectivity in sequential blocks of ten, within and across chains; the last ten neurons are connected to the first ten.

Superficial to deep layer	Corresponding sequential blocks of ten converge onto the deep layer
Deep PFC to MD	All deep layer chains 1&2 converge onto the first block of MD shape (30%), and from 3&4 converge onto the first block of MD orientation (30%)
Deep PFC to VA	From PFC deep first two blocks (40%, chains 1 & 2/ chains 3 & 4) converge to all VA cells (shape/ orientation)
VA to PFC ensemble (driving)	VA to PFC 40% rule selective ensemble
VA to PFC superficial (modulatory)	20% of VA cells projected back to all superficial PFC cells
MD to PFC	All MD to 50% of PFC rule selective ensemble and MD randomly projects to the 10% superficial PFC cells
PFC ensemble to MD	All to the same rule MD

Supplementary Table 2. Model synaptic weights

Synaptic Weights	
PFC middle to PFC superficial	0.02
PFC superficial within chains (except rule ensemble)	0.0012
PFC superficial and deep across chains	0.0009
PFC superficial to deep within chains	0.008 for the first 10 superficial neurons to first 10 deep neurons and 0.004 for the other neurons
PFC superficial to deep across chains	0.0016 for the first 10 superficial neurons to first 10 deep neurons and 0.0008 for the other neurons. As shown in Figure S10, we also increased weights from PFC superficial to deep across chains representing the same abstract rule. See "PFC layers" subsection in "Network structure" section of Methods for details.
PFC deep to MD	0.003
PFC deep to VA	0.015
PFC rule to MD	0.0018
PFC rule within chains	0.002
PFC deep to ST	0.0145
VA to PFC rule	0.01
VA to PV	0.0145
MD to PFC rule	0.0036
MD to PFC superficial	0.012
MD to FS	0.0005

FS to PFC rule	0.0025
ST to SNpr and SNpr to VA	0.001
PV to PFC superficial	0.01

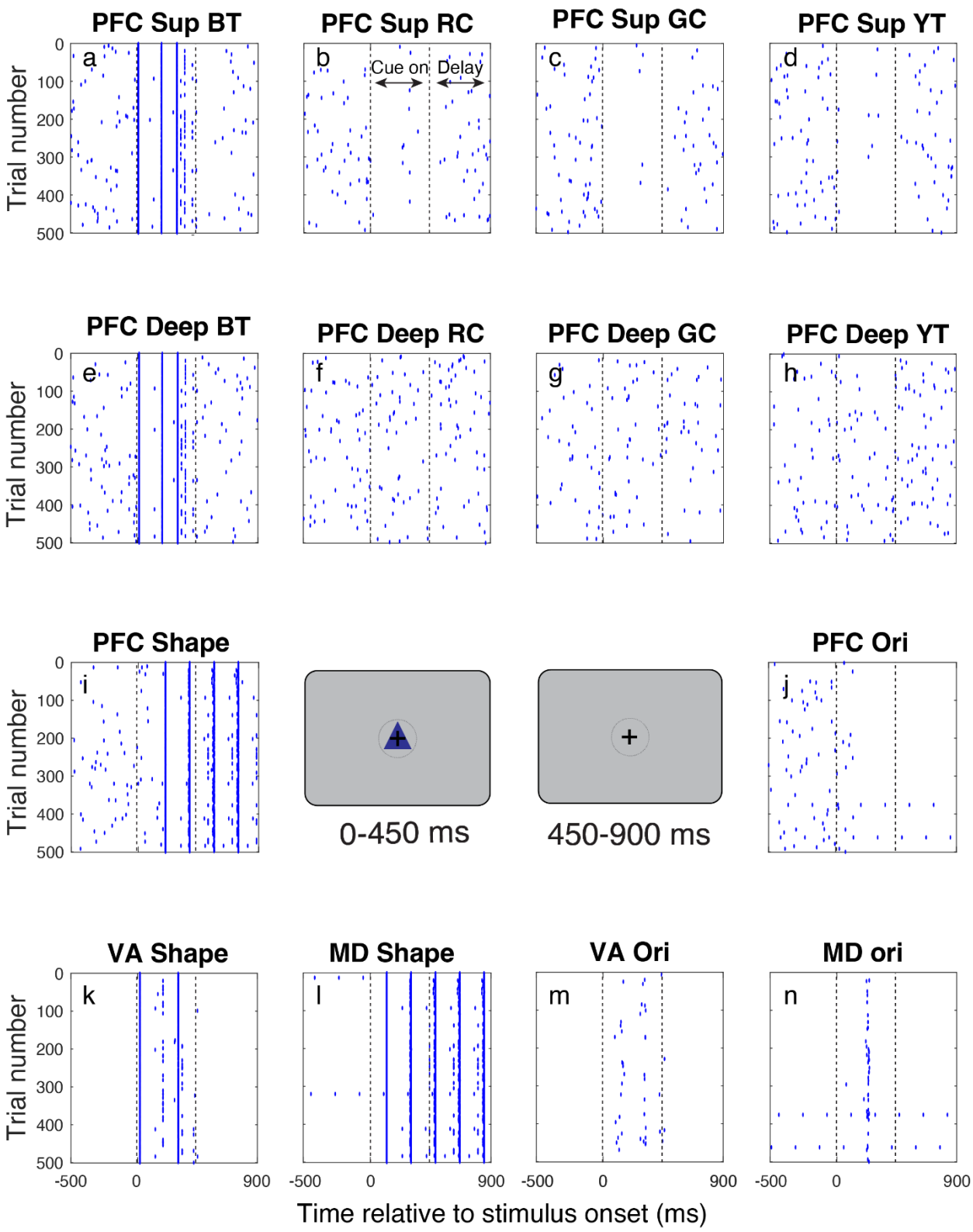
PFC Sup BT
PFC Sup RC
PFC Sup GC
PFC Sup YT
PFC Deep BT
PFC Deep RC
PFC Deep GC
PFC Deep YT
PFC Shape
PFC Ori
VA Shape
MD Shape
VA Ori
MD ori
a
b
c
d
e
f
g
h
i
j
k
l
m
n
Cue on
Delay
Trial number
0
100
200
300
400
500
-500
0
900
0-450 ms
450-900 ms
Time relative to stimulus onset (ms)

Figure S9. *Example model spike output (rasters) for blue triangle cue (shape rule) trials*. Each raster corresponds to individual cell from model populations described in Figure 6a. Insets represent modeled HRT trial epochs: visual stimulus presentation (abstract rule cue) and delay period. Sup, superficial; BT, blue triangle; RC, red circle; YT, yellow triangle; GC, green circle; Ori, orientation.

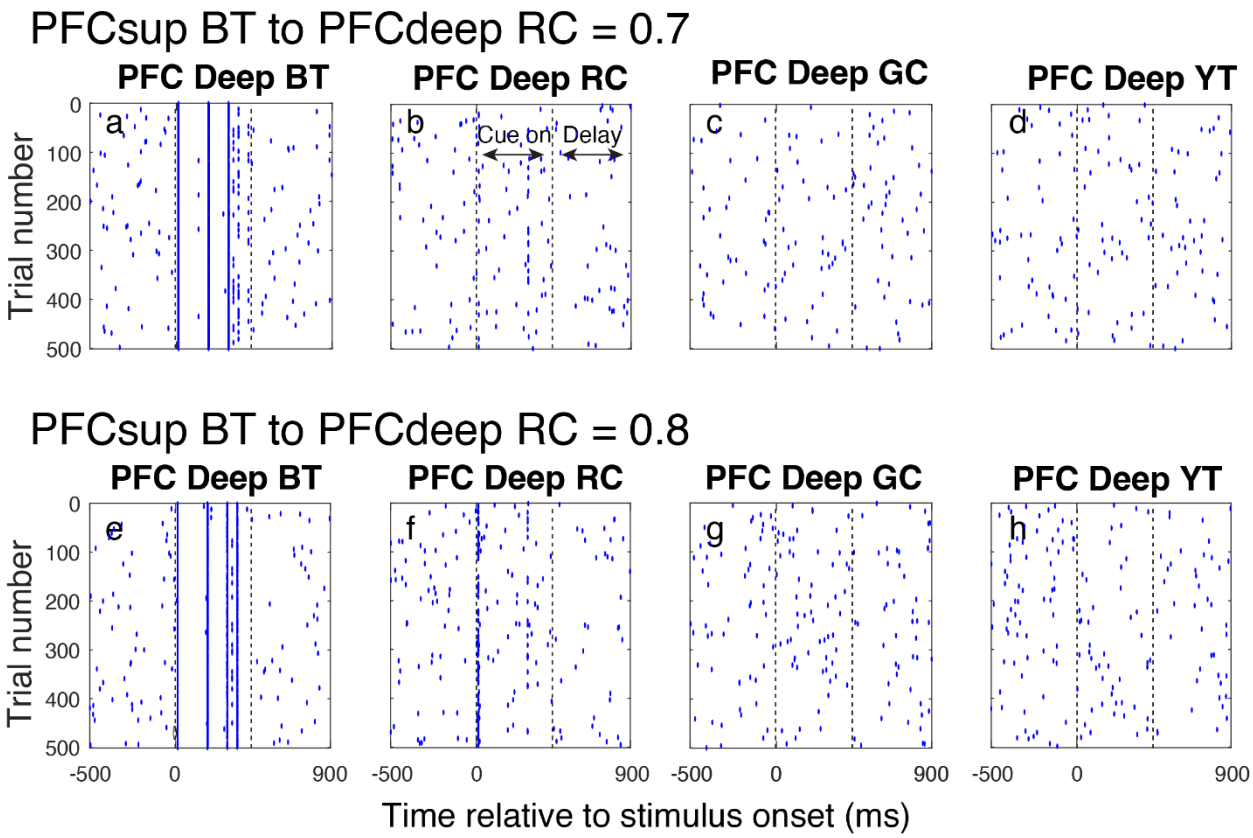

Figure S10. *Increasing convergence from PFC superficial layers to PFC deep layers* in silico *does not generate early abstract rule selectivity in PFC, in line with* in vivo *results*. Model spike output (rasters) for PFC deep layer cells shown for two different model configurations, on blue triangle cue (shape rule) trials. Each row corresponds to a model with a different synaptic weight (convergence) from PFC superficial BT cells to PFC deep RC cells. For comparison, the model configuration in Figure S9 has four populations of PFC superficial cells (BT, RC, GC, YT) converging on to four populations in the deep layer, with a connectivity matrix that has the strongest weight on the diagonal (**2* from the superficial BT population to the deep BT population) and weaker weights off the diagonal (**0.2* from the superficial BT population to deep populations RC, GC and YT). Here, in Figure S10, the model configurations have adjusted weights off the diagonal that correspond to the same rule: (top row) **0.7* from the superficial BT population to the deep RC population; and (bottom row) **0.8* from the superficial BT population to the deep RC population. BT, blue triangle; RC, red circle; GC, green circle; YT, yellow triangle.