Recurrent neural network models of multi-area computation underlying decision-making

Michael Kleinman[a], Chandramouli Chandrasekaran[b,c,d,*], Jonathan C. Kao[a,e,*]

[a]*Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, 90095*

[b]*Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA*

[c]*Department of Anatomy & Neurobiology, Boston University School of Medicine, Boston, MA, USA*

[d]*Center for Systems Neuroscience, Boston University, Boston, MA, USA*

[e]*Neurosciences Program, University of California, Los Angeles, CA, 90095*

Abstract

Cognition emerges from the coordination of computations in multiple brain areas. However, elucidating these coordinated computations within and across brain regions is challenging because intra- and inter-areal connectivity are typically unknown. Testable hypotheses about these interactions are also generally unavailable. To study coordinated computation, we trained multi-area recurrent neural networks (RNNs) to discriminate the dominant color of a checkerboard and output decision variables reflecting a direction decision, a task previously used to investigate decision-related dynamics in dorsal premotor cortex (PMd) of monkeys. We found that multi-area RNNs, as opposed to single RNNs, reproduced the decision-related dynamics observed in PMd during this task. The RNN solved this task by a novel mechanism in which RNN dynamics integrated color information on an axis subsequently readout by an orthogonal direction axis. Direction information was selectively propagated through preferential alignment with the inter-areal connections, while the color information was filtered. These results suggest that cortex uses modular computation to generate minimal but sufficient representations of task information, selectively filtering unnecessary information between areas. Finally, we leverage multi-area RNNs to produce experimentally testable hypotheses for computations that occur within and across multiple brain areas, enabling new insights into distributed computation in neural systems.

Keywords: Decision making, Recurrent Neural Networks (RNNs), Multiple Brain Areas, Subspaces, Modularity, Winner-Take-All.

Introduction

Attention, multisensory integration, motor control, timing and decision-making all emerge from the coordination and cooperation between multiple interconnected brain areas[1–9]. While neural activity

*These authors contributed equally

**Corresponding Authors: kao@seas.ucla.edu and cchandr1@bu.edu

Email addresses: michael.kleinman@ucla.edu (Michael Kleinman), cchandr1@bu.edu (Chandramouli Chandrasekaran), kao@seas.ucla.edu (Jonathan C. Kao)

in a particular area can contain behaviorally relevant signals, such as choices or percepts, it is often unclear if these signals originate within the area or are inherited from upstream brain areas[6]. Therefore, understanding the neural bases of these behaviors necessitates an understanding of the intra- and inter-areal dynamics, that is, how neural activity evolves within and between brain areas[10–17]. However, we currently lack clear hypotheses that describe how distinct brain area dynamics and connectivity relate to computation. To address this gap, we use multi-area recurrent neural networks (RNNs) to derive mechanistic hypotheses and insight into how behaviorally relevant computations emerge within and across multiple brain areas.

Optimized feedforward and recurrent neural networks (RNNs) are emerging tools to model and produce insight into cortical computations associated with visual[18,19], cognitive[20–26], timing[3,27,28], navigation[29], and motor tasks[30–33]. The artificial units of these trained RNNs often exhibit heterogeneous neural responses and neural population dynamics observed in brain areas thought to be be involved in these cognitive and motor tasks[20,21,24,30,31,34]. An important advantage of RNNs is that the recurrent connectivity between artificial units, as well as their activity, are fully observed. It is therefore possible to engineer[35] and reverse engineer[36] these networks using the representation of the internal units and their known recurrent connectivity to gain insight into possible computations in these brain areas[3,20,23]. To date, the primary means to model recorded activity with an RNN is to train an RNN model where all artificial units are connected, which we term a "single area" RNN. These RNNs accept experimenter-designed task inputs to produce behavior-related outputs. If artificial units resemble cortical neurons, these RNNs are subsequently analyzed to propose hypotheses for how a local computation occurs in the brain[3,20,23,31].

Single-area RNNs, however, do not provide insight into how distributed computation occurs across multiple areas. For example, dynamics within multiple cortical areas, including the prefrontal, parietal, and premotor cortex, are implicated in decision-making[7,9,20,37–41]. Inter-areal connections between these and other brain regions communicate (or reject) information between cortical areas[16,42,43]. Currently, several questions remain unanswered. Most broadly, how can multi-area computation support decision-making? More specifically, what are the roles of within-area dynamics and inter-areal connections in mediating distributed computations? How do we develop testable hypotheses and plausible mechanisms for how cortex uses distributed and multi-area computation for decision-making? These questions are fundamentally inaccessible in a single area RNN, since it models local computation in a particular area rather than distinct computations spanning multiple areas.

To answer these questions, we sought to develop new RNN models of decision-making that explicitly incorporate multi-area computation. In particular, we focused on modeling a task where cortex filters task-relevant information prior to output cortical areas, indicating that output and preceding areas perform distinctly different computations. In this 'Checkerboard Task'[37], shown in Fig 1a, the monkey was first shown red and green targets whose location (left or right) was randomized on each trial ('Targets' epoch in Fig. 1a). He was subsequently shown a central static checkerboard composed of red and green squares. The monkey was trained to discriminate the dominant color of the static checkerboard and reach to the target matching the dominant color. Since the target configuration was randomized across trials, this task separates the direction decision from the color decision. Neurons in the caudal portion of dorsal premotor cortex (PMd) covary with the direction, but not color, decision[37,39,44]. This result suggests that the transformation from color to direction decision is computed upstream of PMd, consistent with the view that multiple brain areas are involved in performing this decision task and more generally cognition.

To develop computational models of multi-area computation underlying decision-making with new testable hypotheses, we modeled the Checkerboard Task using RNNs. Single-area RNNs did not resemble PMd

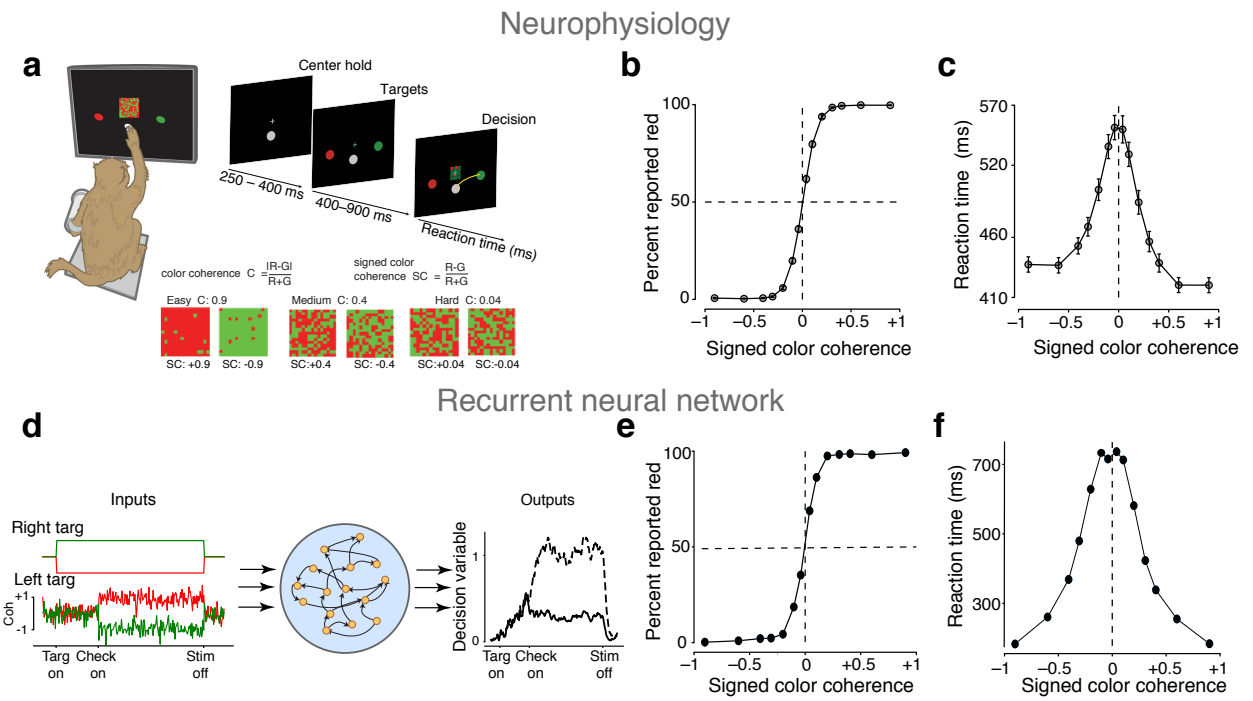

Figure 1: Task design and RNN inputs for a perceptual decision making task. (a) Perceptual decision-making task: A monkey, after holding a center location for $250 - 400$ ms, is presented with a red and green target. The location of these targets (i.e., left or right) is randomly chosen on each trial. After a holding period of $400 - 900$ ms, the monkey is presented a 15×15 checkerboard, composed of red and green squares. To successfully complete the trial, the monkey must reach to the target matching the dominant color of the checkerboard. **(b,c)** Psychometric curve (panel b) and reaction time (panel c) from one monkey from prior experimental work[37]. Figure reproduced from[37] with permission. **(d)** RNN configuration. The RNN receives 4 inputs. The first two inputs indicate the identity of the left and right targets, which can be red or green. These inputs are noiseless. The last two inputs indicate the value of the signed color coherence (proportional to amount of red in checkerboard) and negative signed color coherence (propotional to amount of green in checkerboard). We add independent Gaussian noise to these signals (see Methods). The network outputs two analog decision variables, each of which indicates evidence towards the right target (solid line) or left target (dashed line). A decision is made in the direction of whichever decision variable passes a preset threshold (0.6) first. The time at which the input passes the threshold is defined to be the reaction time. **(e,f)** Psychometric curve (panel e) and reaction time (panel f) for the RNN.

activity. However, multi-area RNNs trained to perform this task resembled PMd both at the level of single unit statistics and neural population dynamics. In particular, the last area of the multi-area RNNs represented the direction decision but not the color decision. We explored the roles of intra-areal dynamics and inter-areal connections in performing this task by investigating the internal representations and connectivity in the RNN. We took advantage of the RNN's full observability to propose a mechanism for (1) computing the direction decision through recurrent dynamics upstream of the output area and (2) selectively propagating direction information through inter-areal connections. Together, our investigation and results provide a demonstration that distributed computation can be understood in terms of intra-areal dynamics that are selectively propagated to subsequent areas. These analyses provide a framework for studying distributed computations in neural systems underlying cognition. We use this framework to develop clear and testable hypotheses for computations spanning brain areas.

Results

We trained monkeys to perform the Checkerboard Task[37] (Fig. 1a). The checkerboard was a static 15×15 grid composed of red and green squares. The number of red and green squares was randomized across trials to create different levels of discrimination difficulty. The signed color coherence linearly indicates the degree to which color is dominant in the checkerboard, with -1 corresponding to completely green, 0 to equal numbers of red and green squares, and $+1$ to completely red. If the monkey reached to the target matching the dominant checkerboard color, the trial was counted as a success and the monkey received a juice reward. Critically, as noted before, the color decision (red or green) is decoupled from the direction decision (left or right) because the left and right target identities are random on every trial. Fig. 1b,c show the monkey's psychometric and reaction time (RT) behavior on this task. Monkeys react slower for more ambiguous checkerboards compared to the almost completely red or completely green checkerboards[37].

We trained RNNs to perform the same Checkerboard Task, as shown in Fig. 1d. The RNN processed four inputs: (1) the color of the left target, (2) the color of the right target, (3) the signed color coherence, which is proportional to the number of red squares, and (4) the negative signed color coherence, which is proportional to the number of green squares. Although the checkerboard input is static, the representation of the checkerboard is not. We therefore added Gaussian noise to the 3rd and 4th inputs, a perturbation to model the noisy inputs also used in a prior RNN study[20]. A red (green) target was denoted by a value of -1 $(+1)$. The outputs of the RNN were two decision variables (DVs), one accumulating evidence for a reach to the right (solid) and the other for a reach to the left (dashed). We defined a choice as when any of the DVs exceeded a preset threshold of 0.6. The reaction time was defined as the first time the DV exceeded this threshold. If the RNN output failed to exceed the threshold by the end of the trial, we counted the RNN's choice as the DV with the higher value.

We used a continuous-time RNN, trained via gradient descent using backpropagation through time and the Adam optimizer[45]. To train these RNNs, we incorporated parameter and rate regularizations, including gradient clipping and regularization to ameliorate vanishing and exploding gradients[46]. Like prior studies[20,21], we terminated training early so that RNNs more faithfully reproduced the monkey's behavior; otherwise, RNNs would achieve higher performance at lower coherences compared to monkeys. RNN psychometric and RT curves were similar to those of monkeys performing this task (compare Fig. 1e,f to Fig. 1b,c). We note that our stopping criterion could induce behavioral variability across trained RNNs. However, we did not observe such a pattern. The psychometric and RT curves of multiple trained RNNs used in this study are shown in Supp. Fig. S1, demonstrating stereotyped behavior.

We next evaluated the activity and dynamics of trained RNNs to assess their similarity to neurophysiological recordings from PMd. Our results are organized as follows. We first show that single area RNNs fail to replicate PMd dynamics in the Checkerboard task, whereas multi-area RNNs more faithfully replicate PMd dynamics. We then investigate the computations in the multi-area RNN that enable it to better reproduce PMd dynamics.

Single area RNNs do not faithfully model PMd neural population activity

In the Checkerboard Task, neural activity in PMd principally reflects the direction decision (left or right) and not the dominant color of the checkerboard (red or green)[37,47]. To summarize this phenomenon, we show in Fig. 2a the principal components (PCs) of the PMd neural population activity. These PC trajectories exhibit separation based on the eventual reach direction (right reaches in solid, left in dotted). These trajectories, however, do not separate based on whether the dominant color was red or green; for

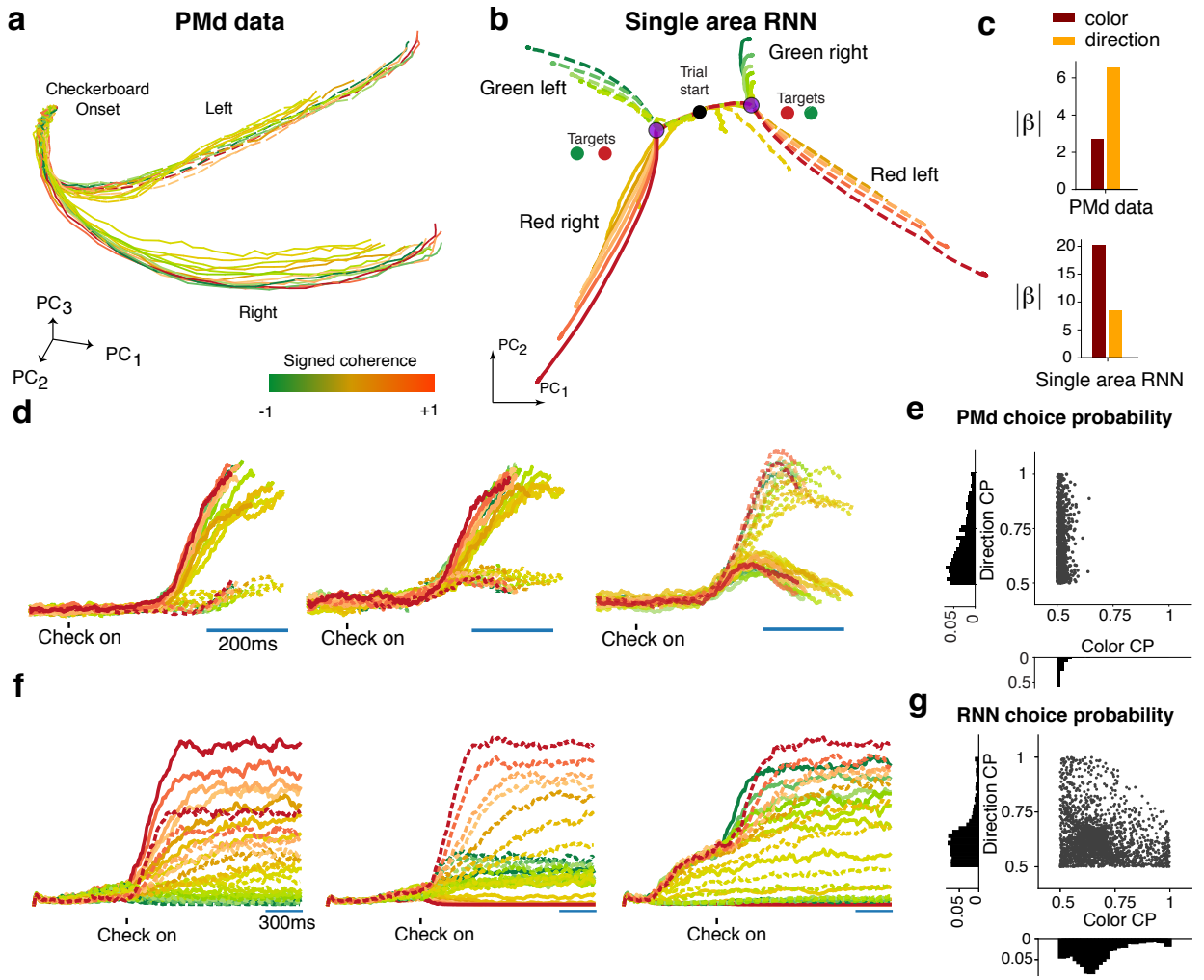

Figure 2: PMd and fully connected RNN dynamics. (a) PMd neural trajectories in the top 3 PCs of the neural population activity. The color reflects the coherence, with darker shades of red (green) indicating more red (green) checkerboards. Solid lines correspond to right reaches, while dotted lines correspond to left reaches. Note the covariation with color coherence but modest covariation with signed color coherence. **(b)** RNN neural trajectories in the top 2 PCs. **(c)** Regression coefficients for the color (maroon) and direction axis (orange) for the PMd data (top) and RNN (bottom). **(d)** Example PMd PSTHs, with activity aligned to checkerboard onset. **(e)** Choice probability for all recorded PMd units. **(f)** Example RNN PSTHs. (Left) Primarily coherence selective, (middle) primarily direction selective, and (right) mixed selectivity. **(g)** Choice probability for all simulated RNN units. For all PSTHs and trajectories, a red-yellow-green colormap is used. The yellow colors have been altered in saturation and brightness to be more visible when printed.

example, the trajectory corresponding to the dominant green checkerboard (signed color coherence -0.9) and dominant red checkerboard (signed color coherence $+0.9$) overlap (Fig. 2a).

We trained a single area RNN to determine if it exhibited similar dynamics. The dynamics of this RNN were different than PMd, with a representative example shown in Fig. 2b. While PMd demonstrated two distinct trajectory motifs predominantly related to the direction decision, the single area RNN demonstrated four trajectory motifs for each combination of (left vs right) and (red vs green). In the Targets epoch, the RNN's activity approaches one of two attractors in state space (purple dots), corresponding to the two target configurations (i.e., red target on left or right). In the checkerboard epoch, trajectories separate based on the coherence of the checkerboard, causing 4 total distinct trajectory

motifs. The single area RNN therefore has a strong color representation and does not faithfully model PMd dynamics.

To quantify this disparity, we used targeted dimensionality reduction[20] (TDR) to identify the dominant axes coding for the signed color coherence ("color axis") and direction decision ("direction axis," see Methods). Because TDR employs linear regression to identify these axes, the magnitude of the regression coefficients for the color and direction axis indicate the degree to which the checkerboard coherence and the direction decision are coded in a neural population. We quantified the magnitude of the coefficients for the color and direction axis to assess how much color and direction information was present in PMd data and the single area RNN. For PMd data, direction related information was the dominant signal (2.7 for color and 6.6 for direction, Fig. 2c, top panel). In contrast, for the single area RNN, color was the dominant signal (20.2 for color and 8.5 for direction, Fig. 2c, bottom). These results show a stronger color representation in the single area RNN, a phenomenon inconsistent with our recorded PMd data.

Differences between PMd and the single RNN were also apparent at the single unit level. PMd peri-stimulus time histograms (PSTHs) reflect reach direction and task difficulty (absolute value of signed coherence) and do not reflect color differences, as shown in Fig. 2d. In contrast, artificial RNN units had a diversity of representations. We found units that primarily reflected checkerboard coherence, Fig. 2f (left); units that reflected the eventual action choice, (middle); and units with a mixed representation, so that color preference could depend on the eventual action choice (right).

We quantified this diversity of responses by calculating the choice probabilities (CPs) reflecting how well direction decision (left or right, direction CP) or color decision (red or green, color CP) could be decoded from single neurons (see Methods). We found, as expected, that PMd neurons had high direction, but low color, choice probability (Fig. 2e). In contrast, across 8 separate RNNs, each with 300 artificial units, we found many single area RNN units had significantly above chance color choice probability (Fig. 2g). In particular, 71.5% (56%) of artificial units had a color (direction) CP greater than 0.6, and further, many neurons had mixed selectivity with above chance color and direction CP. Together, these results demonstrate that a single area RNN is not an appropriate *in silico* model of PMd for this task.

Multi-area RNNs exhibit PMd-resembling neural population dynamics

The brain, like the single area RNN, must make use of color information to solve the checkerboard task. Because PMd activity does not reflect the color decision, we reasoned that the color decision is transformed into a signal that contains the direction decision upstream of PMd. Areas, including the dorsolateral prefrontal cortex (DLPFC), and the ventrolateral prefrontal cortex, have been implicated in such sensorimotor transformations[4,9,48–50].

To abstract various stages of processing, we designed an RNN with three explicit areas by constraining the connectivity of $\mathbf{W}_{\text{rec}}$ into three separate subpopulations of neurons[21]. We consider the three areas to be dorsolateral prefrontal cortex (DLPFC), rostral PMd (PMdr), and caudal PMd (PMd). These are areas that are anatomically connected in a loosely hierarchical manner . DLPFC is connected to PMdr but not PMd and PMdr sends strong projections to PMd[51,52].

An example of a constrained connectivity matrix, $\mathbf{W}_{\text{rec}}$, is illustrated in Fig. 3a. This RNN had varying levels of feedforward and feedback connections based on the proportion of projections between prefrontal and premotor cortex from a macaque atlas[53]. We varied the proportion of feedforward connections between RNNs across values of 10%, 20% and 30% and feedback connections across values of 0%, 5% and 10%. Feedforward and feedback connections are highlighted by yellow and purple squares, respectively, in Fig. 3a. Like the single RNN, we also designated 80% of the units to be excitatory and

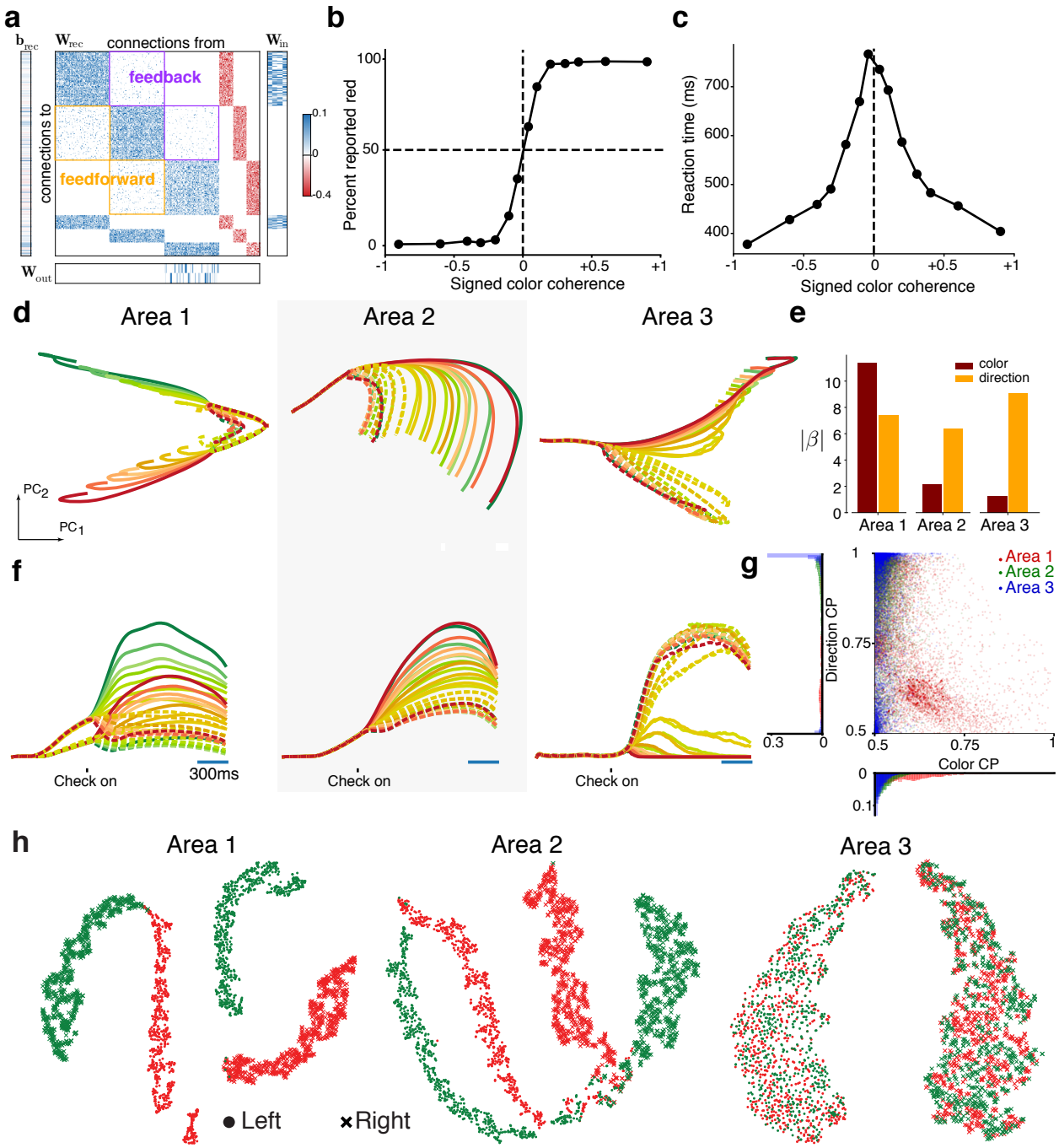

Figure 3: Behavior and artificial unit responses of a multi-area RNN. (a) Example connectivity matrix of RNN with three areas, an $80\%/20\%$ split of excitatory / inhibitory units, and obeying Dale's law, stylized as in[21]. **(b,c)** Example psychometric and RT curves of a multi-area RNN. **(d)** Neural trajectories in the top 2 principal components across three areas. **(e)** Color and direction regression coefficients across all three areas. **(f)** Example PSTHs across all three areas. **(g)** Choice probabilities for units in all areas, demonstrating less color choice probability in later areas of the multi-area RNN. **(h)** tSNE embedding across the three areas. Peri-movement firing rates (see Methods) are embedded in a 2D space and labelled with the selected target color and reach direction. Trials to the left (right) are denoted with a '.' ('x'). Unlike Areas 1 and 2, Area 3 only has two clusters separated based on the direction decision.

20% of the units to be inhibitory, with units obeying Dale's law[21]. Inputs were connected to all units of the first RNN and outputs were read out from excitatory units of the last RNN. We found that multi-area RNNs were capable of performing the task with psychometric and RT curves resembling behavioral data (Fig. 3b,c and Supp. Fig. S1).

To gain preliminary insight into the representations of the multi-area RNN, we examined low dimensional projections of the activity in each area. The multi-area RNN had qualitatively different neural dynamics across areas. Fig. 3d shows PC projections from each of the three areas. Area 1, like the single area RNN, had four distinct trajectory motifs. Activity along PC_1 primarily represented right and left reaches, while activity along PC_2 represented red versus green checkerboards. In contrast, we found Area 2 and 3 had less color selectivity; trajectories of similar color coherence (absolute value of signed color coherence) were highly overlapping. Finally, Area 3 neural population trajectories resembled PMd neural trajectories, separating primarily based on reach direction. We performed dimensionality reduction via t-distributed stochastic neighbor embedding[54] (tSNE; see Methods), in case color selectivity was present in higher PCs. In Area 3, neural population activity separated into only two distinct clusters corresponding to the direction of the reach (Fig. 3h). Additionally, we compared the TDR coefficients and found that color representations decreased in later areas (Color coefficients - Area 1: 11.4, Area 2: 2.1, Area 3: 1.2, summarized in Fig. 3e). On the other hand, direction representations increased in later areas (Direction coefficients: Area 1: 7.4, Area 2: 6.4, Area 3: 9.1, Fig. 3e). We also quantified the variance captured by the axes. These axes, though not optimized to maximize variance, captured significant variability related to discriminating color and direction information, as shown in Tables 2 and 3. Together, the PC projections, tSNE projections, and the trend in the TDR coefficients demonstrate that introducing hierarchical structure into the RNN causes Area 3 to resemble neural population dynamics in PMd.

We also observed this resemblance on the single unit level, with Area 3 PSTHs more closely resembling PMd PSTHs. Example PSTHs from the three areas are shown in Fig. 3f with their respective choice probabilities summarized in Fig. 3g. Area 1, like the single area RNN, represented both direction and color information. The color choice probability progressively decreased in Areas 2 and 3.

Together, this constellation of results shows that Area 3 of the multi-area RNN more faithfully models PMd activity, and is therefore a more realistic representation of how decision-making evolves in the brain during the checkerboard task. Our results also highlight that multi-area RNNs are better than single-area RNNs in modeling neural dynamics when computations span multiple brain areas.

A multi-area RNN uses both recurrence and inter-areal computation to perform the Checkerboard task

We next interrogated the computations in the multi-area RNN. In particular, we were interested in the following question: how does the multi-area RNN transform its inputs so that direction information, but not color information, is preserved in Area 3, like our experimental data in PMd? Understanding the mechanisms of the multi-area RNN's computations reveals how a recurrently connected network uses dynamics and inter-areal connections to not only solve this task, but do so in a way resembling neurophysiology. Analysis of the dynamics in this multi-area model will in turn produce new hypotheses for the computations occurring in areas upstream of PMd, such as the PMdr and DLPFC.

Both the brain and multi-area RNNs use a series of computations that lead to decreased color selectivity in downstream areas. Taking advantage of the full observability of the multi-area RNN, we analyzed the RNN's parameters and activity to study this decrease in color selectivity. The multi-area RNN has two aspects of computation: within-areal and inter-areal computation. Within an area, the RNN processes color and direction information through its inputs and recurrence. Between areas, the RNN utilizes

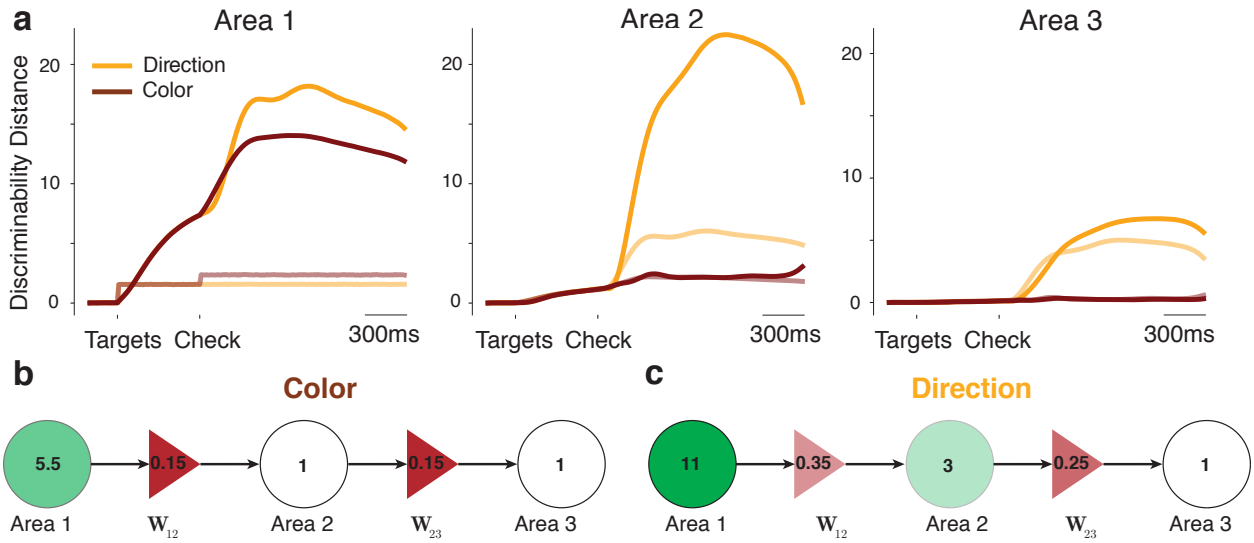

Figure 4: Information discriminability across areas. (a) The norm of the direction discriminability (left red - right red) and color discriminability (left green - left red) as a function of the processing area. The inputs are shown in lighter transparency and the overall activity is shown in solid lines. Area 1 has significant recurrence evidenced by a large separation between the input and overall activity. Between Area 1 and Area 2, there is a larger attenuation in color discrimability (gain 0.15x) relative to direction discriminability (gain 0.35x). Area 2 amplifies the direction discriminability (3x), while the color discriminability stays relatively constant. Between Area 2 and Area 3, there is a larger attenuation in color discriminability (gain 0.15x) versus direction discriminability (gain 0.25x). Area 3 is primarily input driven, with the overall activity being of similar magnitude to the inputs. **(b)** Color and **(c)** direction gain through recurrence (circles) and feedforward connections (triangles), show how color information is filtered throughout the network.

inter-areal connections to selectively propagate information. Hence, we first asked: does the multi-area RNN filter color information through primarily recurrence, inter-areal connections, or a mixture of both?

To disentangle the input contributions from the recurrent contributions, we compared how the inputs and activity were represented in each RNN area (see Methods). We quantified the direction discriminability of each area by subtracting the activity of right reaches from left reaches during reaches to a red target, i.e., "left red - right red." Similarly, we quantified the color discriminability of each area by calculating "right red - right green." Although we only show discriminability under one condition (e.g., for direction, fixing the color to be red; or for color, fixing the direction to be right), our results are consistent no matter the chosen fixed condition (Supp. Fig. S2a) or whether we quantified the recurrent contribution explicitly (Supp. Fig. S2b). Fig. 4a, b shows the magnitude of these differences using the $L2$-norm, which is a measure of Euclidean distance in the high-dimensional space.

Fig. 4a (left) shows considerable color and direction discriminability in Area 1. This direction decision in Area 1 must emerge through the inputs and recurrence of Area 1. In Area 1, the direction discriminability increases after checkerboard onset, even though the checkerboard input has no discernable projection onto this axis (compare activity line vs. input line in Fig. 4a, left). This difference between discriminability of the inputs and the activity suggests an important role for recurrence in Area 1 (an issue we explore further in the next section). We observe a similar pattern for color information. In Area 1, activity associated with color discriminability was amplified relative to the inputs by approximately 5.5 times, whereas direction discriminability was amplified by 11 times. We report these numbers in the summary plot in Fig. 4b (green bubbles).

The excitatory units of Area 1 are mapped through an inter-areal connectivity matrix to the input of Area 2, denoted $\mathbf{W}_{12}$. At the input of Area 2, we observed the color (direction) discriminability was attenuated,

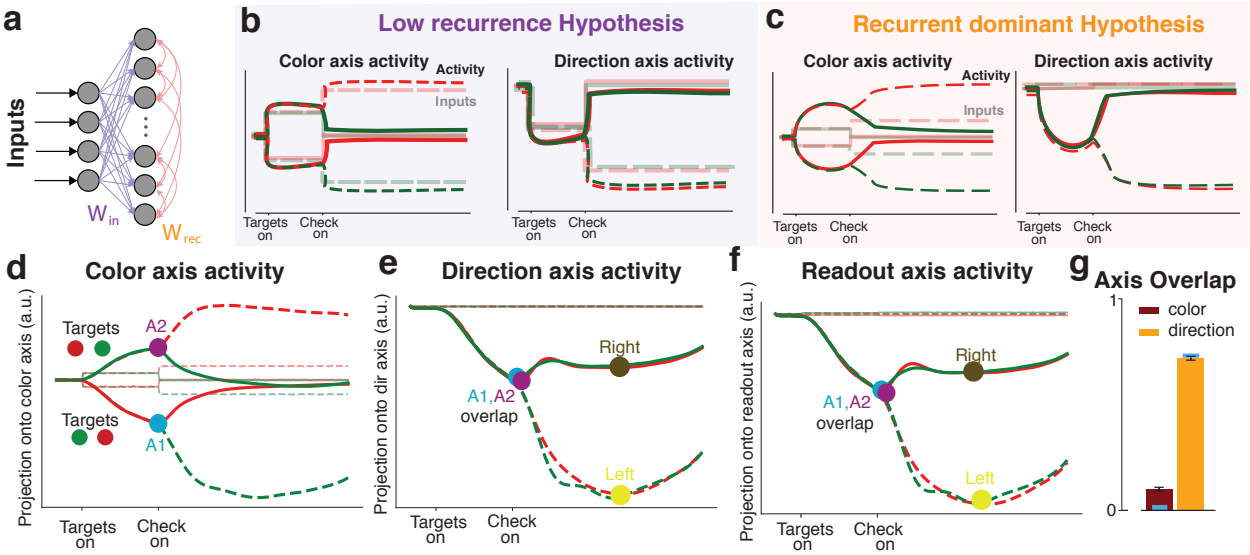

Figure 5: Analysis of Area 1. **(a)** A schematic showing how both inputs and recurrence affect Area 1's activity. **(b)** The low recurrence hypothesis poses that the input activity should be similar to the overall activity and constant through time. **(c)** The recurrent dominant hypothesis suggests a large difference between input and overall activity, without input projections onto the direction axis. **(d)** We projected the activity onto the color and **(e)** direction axis (right panel) identified through TDR. The inputs are shown in the lighter shade and the overall activity is shown in darker shades. The color of the selected target (red or green) is identified by the corresponding color label. Dotted (solid) lines indicate a reach to the left (right). The overall activity initially converges to a target specific attractor (A1 and A2), with the separation in activity occurring based on the color of the checkerboard. Even though there is negligible input projection on the direction axis, the overall activity has a projection onto it. **(f)** Projections onto the top readout axis of $\mathbf{W}_{12}$. **(g)** Overlap of the color and direction axis with the reaodut axis.

having a gain of 0.15 (0.35), indicating that connections between Area 1 and 2 differentially reduced color discriminability. In contrast, the recurrent dynamics in Area 2 amplified direction discriminability by 3 times, but did not amplify color discriminability. Subsequently, inter-areal connections between Area 2 and Area 3, denoted $\mathbf{W}_{23}$, attenuated both color (direction) discriminability further, having a gain of 0.15 (0.25). These results suggest a functional role for Area 2 to selectively amplify direction discriminability, but not color discriminability, before inter-areal attenuation. Finally, there was little difference between color and direction discriminability between Area 3's inputs and activity, suggesting Area 3 is primarily input driven.

Together, these results demonstrate that both within-areal recurrence and inter-areal connectivity play a role in propagating direction information and filtering color information. Further, each area of the RNN appears to play a unique role in the computation. We therefore interrogated these computations further. In particular, we were interested in the following questions. First, how is the direction decision computed? Second, how is color and direction information propagated across areas? Third, what is the mechanism underlying Area 3, which resembles PMd activity? We address these questions in the next sections.

Mechanism for computing the direction decision

The multi-area RNN performs computation across various areas, each one playing a unique role. We first asked how the task is solved, i.e., how is the direction decision computed? Because Area 1 has PC trajectories that separate based on reach direction (Fig. 3d), we reasoned that the direction decision must be computed in Area 1.

We posit two hypotheses for how Area 1 could compute the direction decision. Our first hypothesis, which we term the **"low recurrence"** hypothesis, poses that Area 1 is predominantly input driven. Under this hypothesis, Area 1 responds primarily to current inputs and retains little information about past inputs. This would occur if the recurrent connections in Area 1 play a small role, for instance if only their nonlinearity was used. (We note that it is not possible to solve this task without a nonlinearity, as described in Supplemental Note 1.) The low recurrence hypothesis predicts that: (1) the difference between Area 1 artificial unit activity ($\mathbf{r}_t$) and unit inputs ($\mathbf{W}_{\text{in}}\mathbf{u}_t$) should be relatively constant across time on both a color and direction axis found via TDR, and (2) the inputs should be discriminable on the direction axis. This is diagrammed in Fig. 5b.

Our second hypothesis, which we term the **"recurrent dominant"** hypothesis, poses that Area 1 primarily uses recurrence to reach a direction decision. In this hypothesis, the direction decision unfolds over time through Area 1's dynamics. In contrast to the low recurrence hypothesis, the recurrent dominant hypothesis predicts that: (1) Area 1 artificial unit activity should be significantly larger than unit inputs, with the difference changing through time, (2) unit inputs should have little or no projection on the direction axis, and (3) unit inputs should not be highly aligned with the readout dimensions of Area 1. This is diagrammed in Fig. 5c.

To test these hypotheses, we projected neural population activity on the color and direction axis found via TDR. The color and direction axes are constrained to be orthogonal. We projected the artificial unit activity and unit inputs onto the color and direction axes, shown in Fig. 5d,e. We found that the projection magnitude of the activity was higher than that of the unit inputs on both the color and direction axis. We also observed that the direction axis had negligible projections from the unit inputs, as shown in Fig. 5e. This negligible input projection suggests that the direction axis, representing the action choice, does not directly compute the direction decision from the unit inputs alone. When projecting the activity onto the primary readout dimension of Area 1, the top right singular vector of $\mathbf{W}_{12}$, we found that the unit inputs also had negligible projection on this axis, as shown in Fig. 5f. Further, the readout axis was qualitatively very similar to the direction axis, and we quantified the overlap in Fig. 5g. When quantifying the overlap, we computed the color and direction axis using only the excitatory units, as only excitatory units projected to the subsequent area. Regardless of how we computed the color and direction axes, the projections were very similar (Supp. Fig. S3). Together, these results reject the low recurrence hypothesis for computing the direction decision in favor of the recurrent dominant hypothesis.

Our results suggest that recurrent dynamics in Area 1 are involved in computing the direction decision. We further analyzed the RNN activity and dynamics to discover a dynamical mechanism for computing the direction decision. When considering color axis activity, we observed that target configuration inputs caused a separation of artificial activity to two stable attractors, labeled A1 and A2 in Fig. 5d. However, attractors A1 and A2 overlap on the direction axis, as shown in Fig. 5e. Following checkerboard onset, we observed that color axis activity separated into four distinct trajectories, each for a particular combination of (red vs green) checkerboard and (left vs right) reach. These trajectories are consistent with Fig. 3d, demonstrating that Area 1 represents both the color and direction information. When a red checkerboard was shown, Area 1 dynamics caused activity to increase along the color axis; but when a green checkerboard was shown, dynamics caused activity to decrease along the color axis. Further, the magnitude of left reaches on the color axis was significantly greater than that of right reaches. On the other hand, the direction axis activity did not represent the color choice, but only the direction choice, with two trajectory motifs corresponding to left and right reaches. Like the color axis, the direction axis had larger projection magnitude for left reaches compared to right reaches. When we compared the direction axis with the readout of Area 1 (see Methods), we found a high degree of overlap (0.72, SEM of 0.02), whereas the color axis had a neglible overlap (0.1, SEM of 0.01), with the example RNN

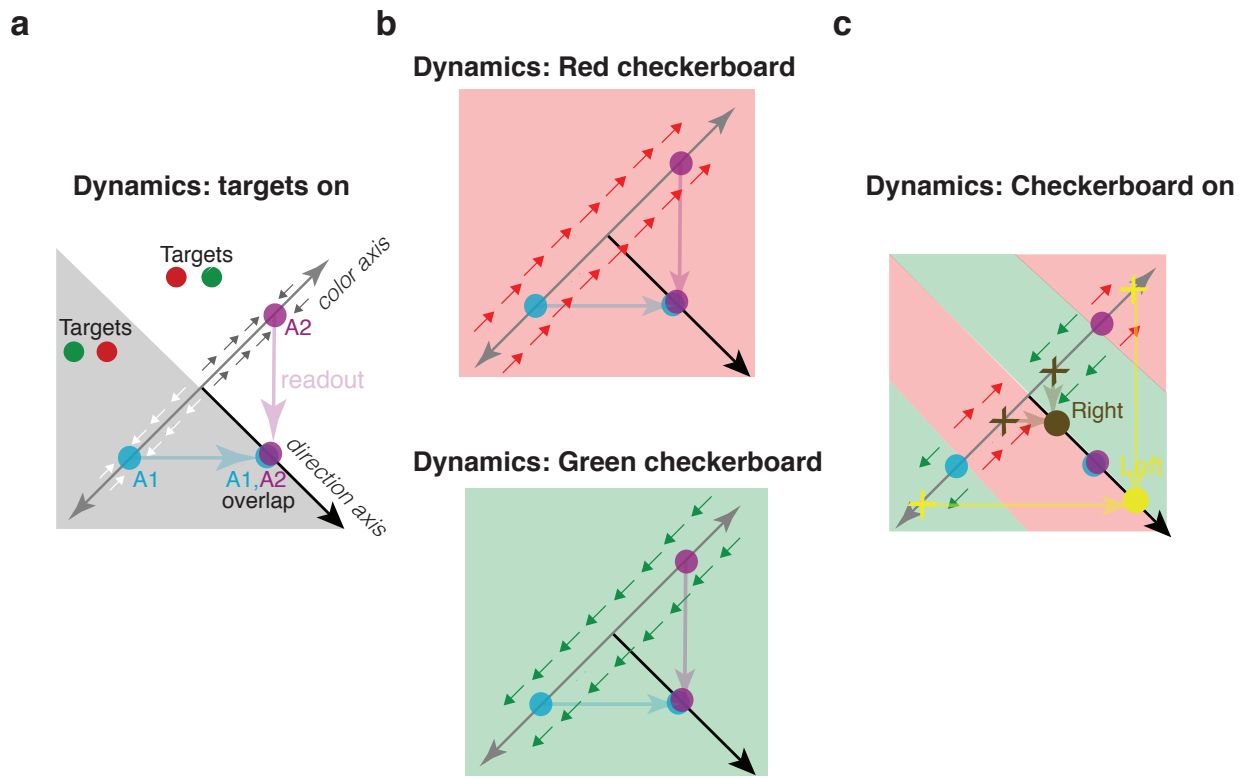

Figure 6: (a-c) A schematic depicting how the overall activity projects onto the color and direction axis, with the target orientation setting the initial condition and movement along the direction axis based on the color of the checkerboard. **(a)** There are two initial conditions on the color axis corresponding to the two different target orientations, but only one corresponding initial condition on the direction axis. **(b)** A red checkerboard causes increasing dynamics along the color axis, while a green checkerboard causes the opposite dynamics. **(c)** Movement along the color axis from either of the two attractors corresponds to a movement along the direction axis towards one of the choices.

highlighted in blue. Together, these results led us to infer that the direction axis is primarily a readout of the color axis activity.

Using these observations, we posit the following mechanism for area 1 of the multi-area RNN, shown in Fig. 6. The color and direction axis are orthogonal, and the direction axis reads out the magnitude of the color axis. We show flow fields consistent with analyzing $\dot{\mathbf{x}}$ in the RNN equations in response to a checkerboard input; these flow fields only apply only to the color axis. When targets turn on (Fig. 6a), dynamics drive the color axis projection to either A1 or A2, depending on the target configuration. However, the readout of the A1 and A2 attractors on the direction axis overlap. When the checkerboard turns on, there are two possible sets of dynamics: red checkerboards cause color axis activity to increase, while green checkerboards cause a decrease (Fig. 6b). Together, these sets of dynamics drive activity along the color axis along four different trajectories following checkerboard onset, but the direction axis readout of the color axis activity only exhibits two trajectories. This is summarized in Fig. 6c.

These results demonstrate a novel mechanism for solving the Checkerboard Task through recurrent dynamics. Even though we identified an axis where the direction information was discriminable, there was a strong color representation within Area 1 along other dimensions. The network must therefore attenuate this color selectivity in later parts of the network. This led us to next investigate the role of inter-areal connections in propagating color and direction information.

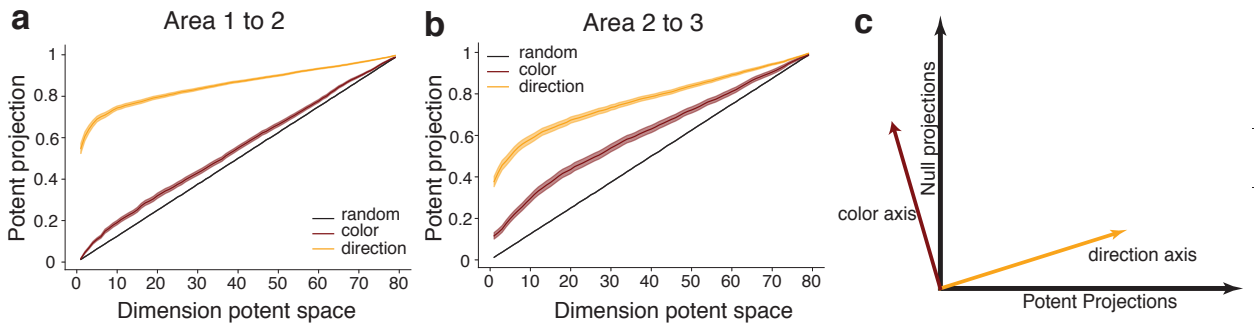

Figure 7: (a) Projections onto the potent space between Areas 1 and 2 for the color and direction axis, and a random vector as a function of effective rank for the input area to the middle area. Regardless of the dimension of the potent space, the direction axis is preferentially aligned with the potent space, indicating the information along this axis propagates, while the color axis is more aligned with the nullspace. Shading indicates the SEM across 72 network instantiations. **(b)** Same as (a) but for projections between Areas 2 and 3. **(c)** Illustration depicting how the orientation of the axes affect the information that propagates.

Inter-areal connections primarily propagate direction and filter color information

As the RNN is fully observed, we have access to the feedforward connections from Area 1 to 2, denoted $\mathbf{W}_{12}$, and from Area 2 to 3, denoted $\mathbf{W}_{23}$. These feedforward connections were constrained to be excitatory connections from excitatory units, as in Fig. 3a. We analyzed the 80×80 matrices of the excitatory inter-areal connections between each area's excitatory units. Based on the hypothesis that the brain uses null and potent spaces to selectively filter information[42], we evaluated the potent and null spaces of $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$. With access to the inter-areal connection matrices, we defined the potent space to be the right singular vectors corresponding to the largest singular values (see Methods). The null space corresponded to the singular vectors with the smallest singular values.

We assessed how the color and direction axis were aligned with these potent and null spaces (see Methods). The projections onto the potent space are shown in Fig. 7a,b for $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$, respectively. The null projection magnitudes are equal to one minus the potent projection, i.e., an inversion of the potent projection magnitudes. These results show that the direction axis is more aligned with the potent space and the color axis is more aligned with the null space. This alignment was robust to the dimension of the effective potent space, and was consistent across all 72 trained networks (SEM shown in lighter shade). This indicates that direction information is preferentially propagated to subsequent areas, while color information is preferentially attenuated. This phenomena is schematized in Fig. 7c.

These results also have implications on how inter-areal connections relay information between areas. Color activity has significant representation in Area 1 (see Fig. 3). Therefore, the inter-areal connections must not merely propagate the highest variance dimensions of a preceding area[16]. We compared the variance captured by the top 2 PCs of Area 1 versus the top 2 readout dimensions of $\mathbf{W}_{12}$ and found that the top 2 readout dimensions of $\mathbf{W}_{12}$ only captured 40% of Area 1's total neural variance. In contrast, the top 2 PCs capture 98% of the variance. We show the variance explained as a function of dimension of the potent space and principal components in Supp. Fig. S4, showing that the dimensions of maximal variance in a preceding area do not necessarily propagate to subsequent areas. Hence, inter-areal connections are not aligned with the most variable dimensions, but are rather aligned to preferentially propagate certain types of information — a result consistent with recent results analyzing activity between V1 and V2[16]. Together, these results provide key insights into how multi-area RNNs selectively filter and propagate particular types of information.

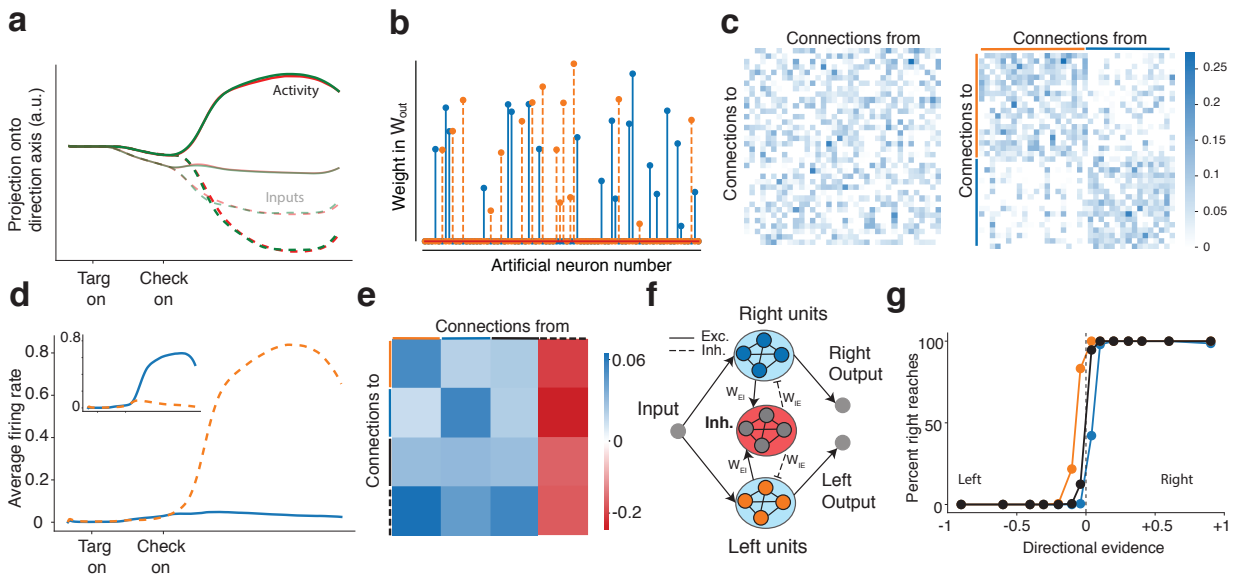

Figure 8: Area 3 mechanism. (a) Projection of input and overall activity onto the direction axis identified through TDR. The conventions are the same as in Figure 5. **(b)** Readout weights in $\mathbf{W}_{\text{out}}$ are sparse, with many zero entries, and selective weights for a left or right reach. **(c)** The unsorted connectivity matrix for the nonzero readout units (left panel), and the sorted connectivity matrix when the matrix was reordered based on the readout weight pools (right). **(d)** Average PSTHs from units for a leftward reach and (inset) rightwards reach. Only one pool is active at a time. **(e)** Averaged recurrent connectivity matrix. **(f)** Schematic of output area. **(g)** Psychometric curve after perturbation experiment, where 7.5% of inhibitory weights to the left pool (orange) and right pool (blue) were increased (doubled). Directional evidence is computed by using the signed coherence and using target configuration to identify the strength of evidence for a left reach and strength of evidence for a right reach.

Area 3, modeling PMd dynamics, is primarily input driven and implements bistable dynamics

We showed previously that Area 3 reproduces PMd's dynamics in Fig. 3d. Our results suggest that the direction decision has been computed before Area 3 and selectively propagated through the RNN's inter-areal connections. We found that the input to Area 3 (through $\mathbf{W}_{23}$) comprises activity that separates based on direction but not color. This activity must be transformed into eventual DV outputs, which accumulate evidence for a left or right reach.

We found that Area 3 was primarily input driven. When comparing the magnitude of the unit inputs, $\mathbf{W}_{23}\mathbf{r}_t^2$, where $\mathbf{r}_t^2$ are the unit activations of Area 2, and $\mathbf{r}_t^3$, the unit activations of Area 3, we found that the magnitude of $\mathbf{r}_t^3$ was close to that of the unit inputs, as shown in Fig. 8a. As previously mentioned, this direction decision has a strong representation in the unit inputs; and yet Area 3 is not simply propagating inputs through a linear transformation. Rather, the PCs of Area 2 and Area 3 have significant qualitative differences, as shown in Fig. 3d, mediated by Area 3's dynamics.

To analyze Area 3's dynamics, we first observed the connectivity of $\mathbf{W}_{\text{out}}$ to identify neural modes corresponding to the left and right DVs. Surprisingly, we found that $\mathbf{W}_{\text{out}}$'s cofficients were sparse, with 44 out of 80 output weights being identically zero. Further, instead of the neural modes being a linear combination of artificial units, we found instead that these modes were clustered subpopulations of artificial units. We illustrate this clustering in Fig. 8b, where units with non-zero coefficients for the left DV (orange) are distinct from those with non-zero coefficients for the right DV (blue). Artificial units either project to the left or right DV outputs, but not both, suggesting that there are two clusters mediating left and right choices.

Based on this clustering, we sorted the units and visualized the excitatory units of $\mathbf{W}_{\mathrm{rec}}$, which upon first glance generally has no discernible structure, as shown in Fig. 8c (left). After sorting, we found that two self-excitatory pools of units emerged in $\mathbf{W}_{\mathrm{rec}}$, the first pool in Fig. 8c (right) corresponding to the left DV and the second pool corresponding to the right DV. The units in each pool have strong excitatory connections, so that, absent other inputs, the pool will tend to self-excite. In addition to these two pools, we identified a pool of randomly connected excitatory units and a pool of inhibitory units with strong projections from and to the two pools. The full connectivity matrix is shown in Supp. Fig. S5. This structure is consistent with a winner-take-all network, where increasing activity in one pool inhibits activity in the other pool through a separate inhibition pool. These observations were consistent with the results obtained by analyzing the pool averaged PSTHs. When one pool increased its activity, the other pool decreased its activity, demonstrating winner-take-all behavior (Fig. 8d).

By taking the averaged connectivity matrix, similar to[25], we confirmed that there were two excitatory pools that received similar projections from the random excitatory pool and inhibitory pool (Fig. 8e). We summarize the behavior with a schematic of the area in Fig. 8f. We subsequently applied selective perturbations to $\mathbf{W}_{\mathrm{rec}}$ to determine how behavioral performance was biased. We increased inhibition to either the right or the left pool by doubling the weights of 7.5% of the inhibitory neurons. We found that this biased the network towards more left or right reaches, respectively, shown in Fig. 8g. When inhibition was increased to the right excitatory pool, the network was more likely to respond left. Conversely, when inhibition was increased to the left excitatory pool, the network was more likely to respond right.

Together, these results show that the output area, modeling PMd, robustly transforms separable inputs to a decision variable through a winner-take-all mechanism. These results also highlight that Area 3 is not just a readout of an upstream area. Recurrent dynamics in Area 3 sculpt the ultimate direction decision of the network. More generally, our RNN results are consistent with the view that decision-making emerges from a distributed interconnected process[4,6,8,55].

Discussion

Systems neuroscience has transitioned its focus from the study of single-neuron responses to how local populations of neurons coordinate to perform a computation[56–58]. While neural population approaches have led to important insights in motor control[42,59–63] and decision making[20,38,64–66], there has been limited insight into how interacting regions, together, coordinate to produce cognitive behavior[16]. In this work, our goal was to address this gap by developing models that provide insight into how computations are coordinated within and across areas to drive behavior.

To this end, we trained single and multi-area RNNs to perform a decision-making task previously investigated in monkeys. We found that the output area of the multi-area RNN did not represent color information but encoded direction decisions, consistent with experimental data in PMd. A single area RNN in contrast, showed both direction and color selectivity. Our multi-area RNN model allowed us to mechanistically interrogate how recurrent dynamics in different areas and inter-areal connections coordinate to perform a decision-making task. Because the parameters of a trained RNN are fully observed, we were able to "open the black box"[34,36], probing how computation evolves in local areas and how information selectively propagates to subsequent areas. Our results are relevant for modeling studies that use RNNs to propose nonlinear dynamical mechanisms in neural circuits, develop theories of inter-areal information processing, and produce testable hypotheses for current and future animal experiments. In the rest of the discussion, we situate and discuss the relevance of our results for these studies.

RNNs have provided insight into neural dynamical mechanisms in the lateral intraparietal area[21,23], prefrontal cortex[3,20], and motor cortex[30–32]. Barring a few exceptions, these RNNs have only been used to model a local computation[7,67]. Our results show that this approach only provides partial insight into neural computations, especially when the computation is distributed across areas. For the Checkerboard Task, we found that single area RNNs do not reproduce dynamics observed in PMd during a decision-making task, but the output area of the multi-area RNNs did. This sugests that the inputs to PMd differ from the inputs to our model in the single area RNN, whereas inputs to the output area are more faithful. We emphasize that we did not constrain RNN training to reproduce PMd dynamics or filter color information. Rather, resemblance emerged through the introduction of multiple areas into the RNN architecture, emphasizing the importance of incorporating architectural features to more faithfully model neural computation. It is also worth noting that only the output area of the RNN resembled PMd, but not any areas before it, as may be expected because PMd is close to the output from the cerebral cortex, sending projections down the spinal cord and to primary motor cortex. This success in replicating experimental observations in PMd highlights the utility of incorporating architectural constraints in RNN modeling. Applying our multi-area RNN to datasets that include electrophysiological recordings from multiple brain regions will provide additional insights into how multiple areas coordinate to drive behavior[4,7,8,17].

Our results suggest that cortex and multi-area RNNs may share a more general principle of multi-area information processing: if information becomes irrelevant for later computations, it is reduced. In the Checkerboard Task, color information is necessary to compute the direction decision. However, color information is unnecessary *after* the direction decision is computed, since the decision is reported only via a motor output. PMd does not represent color information[4,37,39,44,68], indicating that the direction decision is computed in upstream cortical areas such as prefrontal cortex, an area known to represent color information in decision making tasks. By analyzing each RNN area, we also observed that the multi-area RNN computes the direction decision upstream of the output area. In both PMd and the RNN output area, color information was unnecessary because it was used in an upstream area to compute the direction decision. Further, representing the color decision after the direction decision is computed may be disadvantageous in at least three ways: (1) the color decision has zero mutual information with the RNN direction output, (2) representing a redundant color decision may increase the capacity demands of the network, and (3) representing a mixture of color information with a direction decision may obstruct the direction decision representation.

These results appear to be consistent with recent theoretical work interrogating how "information bottlenecks" produce efficient representations in deep (feedforward) neural networks[69,70]. In the information bottleneck framework, there is a tradeoff between having a sufficient and minimal representation of the input for producing the output[71]. When applied to feedforward networks, the information bottleneck framework posits that the performance of neural networks depends on how sequentially connected areas (or layers) transform inputs into more efficient representations for computing the output. New theoretical results suggest that stochastic gradient descent, or a noisy learning process, implicitly optimizes the information bottleneck in neural networks[70]. In our results, we found the direction decision was computed by the input area, providing a sufficient representation for the output. The presence of a directional input for subsequent areas enabled compression of the representation in these areas by discarding the no longer necessary color information. We emphasize that we did not constrain the RNN to achieve this minimal but sufficient representation. Rather, the emergence of this representation was tied to incorporating multiple areas into the RNN, with a feedforward connectivity bottleneck between areas. Our observation that outputs of the multi-area RNN (output area) and cortex (PMd) both discard color information suggest that the brain may also be optimized to produce representations optimized to an information bottleneck, using both sufficient and minimal representations. As later described, this produces testable

hypotheses for future neurophysiological experiments.

We chose 3 areas in our multi-area RNN. We want to emphasize that we are in no way suggesting that 3 areas are ideal or necessary for solving the task. Based on the information bottleneck theory, we predict that a successfully trained multi-area RNN, with an arbitrary number of areas, should exhibit an attenuation of color information in later areas. Our 3 area RNN should be viewed as a demonstration of the importance of incorporating multiple areas in an RNN, and as a simple, base case for gaining insight into inter-areal computation. Ultimately, the number of areas used for modelling neurophysiological tasks should depend on experiments[4,7,8] and anatomical studies[51,52].

A multi-area RNN can perform computation through recurrence and inter-areal connections. However, it need not rely on both; for example, inter-areal connections are not necessary to perform the task, since a single area RNN can solve the Checkerboard Task. We found that the multi-area RNN meaningfully used both recurrent computation and inter-areal connections to amplify and attenuate discriminability information in the network, as shown in Fig. 4. We observed that recurrent areas did not attenuate activity; rather, they tended to selectively amplify color and direction discriminability. In some cases, as demonstrated in area 1, this amplification reflects integration of color information: green (red) checkerboards cause activity to increase (decrease) on the color axis, leading to differential amplification in color selectivity.

On the other hand, inter-areal connections appeared to only attenuate activity. Attenuation appears to be related to the specific orientation of inter-areal connections with particular readout dimensions. The inter-areal connections between area 1 and 2 are highly aligned with the direction axis, suggesting its primary role in propagating direction, but not other, information. As such, information not aligned with the direction axis lies primarily in the *null space* of the inter-area connectivity matrix, and are attenuated. This role for inter-areal connections is related to null and potent spaces[42,72] and communication subspaces[16]. Null and potent spaces were proposed as a means of filtering information between areas, while communication subspaces extended this idea to highlight that inter-areal connections do not always align with the largest dimensions of variability in the antecedent area. While these neurophysiological studies could not measure the inter-areal connections, but observed these phenomena by measuring activity in multiple areas during a task, an RNN model enables us to precisely quantify the alignment of inter-areal connections with neural population activity. By explicitly measuring the inter-areal singular vectors, we were able to verify its primary role in an RNN in propagating dimensions aligned with direction information, and that it is a specific readout that doesn't merely propagate the top principal components of neural population activity.

An important advantage of RNN modeling is that probing an RNN enables mechanism discovery and produces testable hypotheses for future experiments. By analyzing the RNN's activations and parameters, we were able to discover a mechanism for computing a sensorimotor direction decision in the Checkerboard Task, investigate the role of inter-areal connections, and propose how an input driven output area produces PMd resembling dynamics. Our analysis of the multi-area RNN leads to several testable hypotheses for future experiments.

A dynamical mechanism in upstream areas: Our results hypothesize that areas upstream of PMd should exhibit mixed selectivity dynamics for color and direction information. However, our RNN makes more specific predictions as to the mechanism in these upstream areas. We anticipate observing neural trajectories that resemble Fig. 3d rather than Fig. 2b. Our hypothesis for this dynamical mechanism parcellates into concrete predictions: (1) neural population activity should diverge to two stable attractors (analogous to an "optimal subspace") based on the target configuration; (2) dynamics should be consistent along a color axis for red and green checkerboards, no matter the target configuration, as in

Fig. 6 and (3) activity along a direction axis, as a readout of color information, should exhibit trajectories with similar dynamics to the color axis activity. We anticipate testing these predictions using recordings from dorsolateral prefrontal cortex (DLPFC) and posterior parietal cortex.

Role of inter-areal connections: The RNN predicts that inter-areal connections read out the direction axis. Experiments that record from multiple areas, such as PMdr, PMd, and DLPFC[4], can test this hypothesis by predicting activity in PMdr and PMdc from different subspaces in DLPFC. Our results hypothesize that direction-related axes in DLPFC should be more predictive of activity in PMdr and PMd. Further, these direction-related axes should not account for the majority of the neural variance in areas computing the direction decision, since the mechanism primarily relies on integrating color information.

Role of inhibitory connections: In the output area of our RNN, our connectivity matrices indicate that inhibitory connections play an important role in mutually inhibiting clusters of neurons that accumulate evidence for a direction decision. When we selectively increased the inhibition to a cluster, we observed that the opposite cluster was more likely to win the competition. We hypothesize that injecting muscimol in areas of PMd should selectively bias the monkey towards one or the other direction of reaching. Similarly, in tasks that involve bilateral brain networks, muscimol in one hemisphere should bias behavior towards the contralateral side, a prediction consistent with experimental results in rodents[73].

Methods

Somatomotor reaction time visual discrimination task and recordings from PMd:

The task, training and electrophysiological methods used to collect the data used here have been described previously[37] and are reviewed briefly below.

Task: Two trained monkeys (Ti and Ol) performed a visual reaction time discrimination task[37]. The monkeys were trained to discriminate the dominant color in a central static checkerboard composed of red and green squares and report their decision with an arm movement. If the monkey correctly reached to and touched the target that matched the dominant color in the checkerboard, they were rewarded with a drop of juice. This task is a reaction time task, so that monkeys initiated their action as soon as they felt they had sufficient evidence to make a decision. On a trial-by-trial basis, we varied the signed color coherence of the checkerboard, defined as $(R-G)/(R+G)$, where R is the number of red squares and G the number of green squares. The color coherence value for each trial was chosen uniformly at random from 14 different values arranged symmetrically from 90% red to 90% green. Reach targets were located to the left and right of the checkerboard. The target configuration (left red, right green; or left green, right red) was randomly selected on each trial. Both monkeys demonstrated qualitatively similar psychometric and reaction-time behavior. Results in Fig. 1b,c correspond to Monkey Ti.

For analyses in Fig. 8f, we combined signed coherence with the target configuration to estimate a parameter we term directional evidence. For instance, a completely red checkerboard with a signed coherence of 1 and a red target on left and green target on right would be considered to have -1 directional evidence. Conversely, if the target configuration was such that the right target was red and the left target was green, the directional evidence would be 1.

Recordings: In our original study, we reported the activity of 996 units recorded from Ti (n=546), and Ol (n=450) while they performed the task[37]. PMd units demonstrate temporal complexity in their firing rate profiles. In[37], we applied the visuomotor index of that measures the degree of sustained activity to separate this population into the broad categories of increased, perimovement and decreased units. Increased units exhibited ramp-like increases in average firing rate, time-locked to the onset of the visual

stimulus, with slopes that varied with stimulus coherence. These are the classic candidates for neurons that might carry an integrated decision variable for the discrimination task Monkey Ol and Ti's PMd units both had low choice color probability and PCA characteristics. Reported analyses from PMd data use units pooled across Monkey Ol and Ti.

RNN description and training

We trained a continuous-time RNN to perform the checkerboard task. The RNN is composed of N artificial neurons (or units) that receive input from N_{in} time-varying inputs $\mathbf{u}(t)$ and produce N_{out} time-varying outputs $\mathbf{z}(t)$. The RNN defines a network state, denoted by $\mathbf{x}(t) \in \mathbb{R}^N$; the ith element of $\mathbf{x}(t)$ is a scalar describing the "currents" of the ith artificial neuron. The network state is transformed into the artificial neuron firing rates (or network rates) through the transformation:

$$\mathbf{r}(t) = f(\mathbf{x}(t)), \tag{1}$$

where $f(\cdot)$ is an activation function applied elementwise to $\mathbf{x}(t)$. The activation function is typically nonlinear, endowing the RNN with nonlinear dynamics and expressive modeling capacity[74]. In this work, we use $f(x) = \max(x, 0)$, also known as the rectified linear unit, i.e., $f(x) = \text{relu}(x)$. In the absence of noise, the continuous time RNN is described by the equation

$$\tau \dot{\mathbf{x}}(t) = -\mathbf{x}(t) + \mathbf{W}_{\text{rec}}\mathbf{r}(t) + \mathbf{W}_{\text{in}}\mathbf{u}(t) + \mathbf{b}_{\text{rec}}, \tag{2}$$

where τ is a time-constant of the network, $\mathbf{W}_{\text{rec}} \in \mathbb{R}^{N\times N}$ defines how the artificial neurons are recurrently connected, $\mathbf{b}_{\text{rec}} \in \mathbb{R}^N$ defines a constant bias, and $\mathbf{W}_{\text{in}} \in \mathbb{R}^{N\times N_{in}}$ maps the RNN's inputs onto each artificial neuron. The output of the network is given by a linear readout of the network rates, i.e.,

$$\mathbf{z}(t) = \mathbf{W}_{\text{out}}\mathbf{r}(t), \tag{3}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{N_{\text{out}}\times N}$ maps the network rates onto the network outputs.

We trained an RNN to perform the checkerboard task as follows. For all networks, we used $N_{\text{in}} = 4$, $N = 300$, and $N_{\text{out}} = 2$. The four inputs were defined as:

1. Whether the left target is red (-1) or green (+1).
2. Whether the right target is red (-1) or green (+1).
3. Signed coherence of red (ranging from -1 to 1), $(R - G)/(R + G)$.
4. Signed coherence of green (ranging from -1 to 1), $(G - R)/(R + G)$. Note that, prior to the addition of noise, the sum of the signed coherence of red and green is zero.

The inputs, $\mathbf{u}(t) \in \mathbb{R}^4$, were defined at each time step, t, in distinct epochs. In the 'Center Hold' epoch, which lasted for a time drawn from distribution $\mathcal{N}(200\ \text{ms}, 50^2\ \text{ms}^2)$, all inputs were set to zero. Subsequently, during the 'Targets' epoch, which lasted for a time drawn from distribution $\mathcal{U}[600\ \text{ms}, 1000\ \text{ms}]$, the colors of the left and right target were input to the network. These inputs were noiseless, as illustrated in Fig. 1d, to reflect that target information is typically unambiguous in our experiment and unlike the checkerboard is not a mix of red and green squares. Following the 'Targets' epoch, the signed red and green coherences were input into the network during the 'Decision' epoch. This epoch lasted for 1500 ms. We added zero mean independent Gaussian noise to these inputs, with standard deviation equal to 5% of the range of the input, i.e., the noise was drawn from $\mathcal{N}(0, 0.1^2)$. At every time point, we drew independent noise samples and added the noise to the signed red and green

coherence inputs. Following the 'Decision' epoch, there was a 'Stimulus Off' epoch, where the inputs were all turned to 0.

The two outputs, $\mathbf{z}(t) \in \mathbb{R}^2$ were defined as:

1. Decision variable for a left reach.
2. Decision variable for a right reach.

We defined a desired output, $\mathbf{z}_{\text{des}}(t)$, which was 0 in the 'Center Hold' and 'Targets' epochs. During the 'Decision' epoch, $\mathbf{z}_{\text{des}}(t) = 1$. In the 'Stimulus Off' epoch, $\mathbf{z}_{\text{des}}(t) = 0$. In RNN training, we penalized output reconstruction using a mean-squared error loss,

$$\mathcal{L}_{\text{mse}} = \sum_{t \in \mathcal{T}} \left| \mathbf{z}(t) - \mathbf{z}_{\text{des}}(t) \right|^2 . \tag{4}$$

The set $\mathcal{T}$ included all times from all epochs except for the first 200 ms of the 'Decision' epoch from the loss. We excluded this time to avoid penalizing the output for not immediately changing its value (i.e., stepping from 0 to 1) in the 'Decision' epoch. Decision variables are believed to reflect a gradual process consistent with non-instantaneous integration of evidence, e.g., as in drift-diffusion style models, rather than one that steps immediately to a given output.

To train the RNN, we minimized the loss function:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{in}} \left\| \mathbf{W}_{\text{in}} \right\|_F + \lambda_{\text{rec}} \left\| \mathbf{W}_{\text{rec}} \right\|_F + \lambda_{\text{out}} \left\| \mathbf{W}_{\text{out}} \right\|_F + \lambda_r \sum_t \left\| \mathbf{r}(t) \right\|^2 + \lambda_\Omega \mathcal{L}_\Omega, \tag{5}$$

where

- $\left\| \mathbf{A} \right\|_F$ denotes the Frobenius norm of matrix $\mathbf{A}$
- $\lambda_{\text{in}} = \lambda_{\text{rec}} = \lambda_{\text{out}} = \lambda_r = 1$ to penalize larger weights as well as rates[31,32].
- $\lambda_\Omega = 2$
- $\mathcal{L}_\Omega$ is a regularization term that ameliorates vanishing gradients proposed and described in prior literature[21,46].

During the training process, we also incorporated gradient clipping to prevent exploding gradients[46]. Training was performed using stochastic gradient descent, with gradients calculated using backpropagation through time. For gradient descent, we used the Adam optimizer, which is a first order optimizer incorporating adaptive gradients and momentum[45].

We stopped training early to prevent the network from overlearning the task. The RNN will otherwise perform the task with better performance than the animal. Prior studies modeling cognitive tasks have also employed early stopping to reproduce animal behavior[20,21]. Every 200 or 500 training epochs, we generated 2800 cross-validation trials, 100 for each of the 28 possible conditions (14 coherences $\times$ 2 target configurations). For each trial, there was a correct response (left or right) based on the target configuration and checkerboard coherence. The RNN's decision was determined to be either: (1) whichever DV output (for left or right) first crossed a pre-set threshold of 0.6, or (2) if no DV output crossed the pre-set threshold of 0.6 by the end of the 'Decision epoch,' then the decision was for whichever DV had a higher value at the end of this epoch. If the RNN's decision on a single trial was the same as the correct response, we labeled this trial 'correct.' Otherwise, it was incorrect. Training was terminated when the proportion of cross-validated correct trials exceeded 0.65. Finally, in training only, we introduced 'catch' trials 10% of the time. On 50% of catch trials, no inputs were shown to the

RNN and $\mathbf{z}_{\text{des}}(t) = 0$ for all t. On the remaining 50% of catch trials, the targets were shown to the RNN, but no coherence information was shown; likewise, $\mathbf{z}_{\text{des}}(t) = 0$ for all t on these catch trials. For single area RNNs, we trained 8 different networks at different seeds.

Multi-area RNN training

We trained the multi-area RNNs by constraining the recurrent weight matrix $\mathbf{W}_{\text{rec}}$ to have connections between the first and second areas and the second and third areas. The amount of feedforward and feedback connections were guided by[75]. Treating the three areas as Area 9-46d (DLPFC), PMDr, and PMDc, we varied the probability of a feedforward connection across levels 0.1, 0.2, and 0.3, with feedback connections at 0, 0.05, and 0.1. Each area had 100 units. Of these 100 units, 80 were excitatory and 20 were inhibitory. Excitatory units were constrained to have only positive outgoing weights, while inhibitory units were constrained to have only negative outgoing weights. Connections between areas were only between excitatory neurons. We used the `pycog` repository[21] to implement these architecture constraints. We trained multi-area RNNs with each of the nine possible feedforward and feedback proportions, each at 8 different seeds. This resulted in a total of 72 trained networks. The psychometric and reaction-time curves of all these 72 trained networks are shown in Supp. Fig. S1.

In the manuscript, results are sometimes shown for one multi-area RNN, while statistics are reported across all multi-area RNNs. When showing results for one multi-area RNN, we used an RNN with feedforward connection proportion 0.1 and feedback connection proportion 0.05. When reporting statistics across multi-area RNNs in the paper, all 72 networks were analyzed.

RNN behavior

To evaluate the RNN's psychometric curve and reaction-time behavior, we generated 200 trials for each of the 28 conditions, producing 400 trials for each signed coherence. For these trials, we calculated the proportion of red decisions by the RNN. This corresponds to all trials where the DV output for the red target first crossed the preset threshold of 0.6; or, if no DV output crossed the threshold of 0.6, if the DV corresponding to the red target exceeded that corresponding to the green target. The reaction time was defined to be the time between checkerboard onset to the first time a DV output exceeded the preset threshold of 0.6. If the DV output never exceeded a threshold of 0.6, in the reported results, we did not calculate a RT for this trial.

Choice probability

To calculate the choice probability for a single unit, we first computed the average firing rate in a window from [-500 ms, $+500$ ms] around the reaction time for each trial. We used the average firing rates calculated across many trials to create a firing rate distribution based on either the color decision (trials corresponding to a red or green choice) or the direction decision (trials corresponding to a left or right choice).

To compute the color choice probability, we constructed the firing rate distributions corresponding to a green choice or red choice. If these two distributions are non-overlapping, then the neuron has a color choice probability of 1; the average firing rate will either overlap with the red or green firing rate distributions, but not both. On the other hand, if the two distributions are completely overlapping, then the neuron has a color choice probability of 0.5; knowing the firing rate of the neuron provides no information on whether it arose from the red or green firing rate distribution. When there is partial overlap between these two distributions, then firing rates where the distributions overlap are ambiguous.

We computed choice probability as the area under the probability density function at locations when the two distributions did not overlap, divided by 2 (to normalize the probability). To calculate the direction choice probability, we repeated the same calculation using firing rate distributions corresponding to a left choice or right choice.

Targeted Dimensionality Reduction

Targeted dimensionality reduction (TDR) is a dimensionality reduction technique that finds low dimensional projections that have meaningful task interpretations. We applied TDR as described by the study by Mante, Sussillo and colleagues[20]. We first z-scored the firing rates of each of the 300 units across time and trials, so that the firing rates had zero mean and unit standard deviation. We then expressed this z-scored firing rate as a function of task parameters using linear regression,

$$r_{i,t}(k) = \beta^1_{i,t}\mathrm{color}(k) + \beta^2_{i,t}\mathrm{direction}(k) + \beta^3_{i,t}\mathrm{context}(k). \tag{6}$$

Here, $r_{i,t}(k)$ refers to the firing rate of unit i at time t on trial k. The total number of trials is N_{trials}. This regression identifies coefficients $\beta^m_{i,t}$ that multiply the mth task parameter to explain $r_{i,t}(k)$. We defined the task parameters as follows:

- $\mathrm{color}(k)$ was the signed coherence of the checkerboard on trial k, given by $(R-G)/(R+G)$.
- $\mathrm{direction}(k)$ was -1 for a left decision and $+1$ for a right decision.
- $\mathrm{context}(k)$ was the target orientation, taking on -1 if the green (red) target was on the left (right) and $+1$ if the green (red) target was on the right (left).

We did not fit a bias term since the rates were z-scored and therefore zero mean. For each unit, i, we formed a matrix $\mathbf{F}_i$ having dimensions $N_{\text{trials}} \times 3$, where each row consisted of $[\mathrm{color}(k),\ \mathrm{direction}(k),\ \mathrm{context}(k)]$. We define $\mathbf{r}_{i,t}$ to be the rate of unit i at time t across all trials. We then solved for the coefficients, denoted by $\boldsymbol{\beta}_{i,t} = [\beta^1_{i,t},\ \beta^2_{i,t},\ \beta^3_{i,t}]^T$, using least squares,

$$\boldsymbol{\beta}_{i,t} = (\mathbf{F}_i^T\mathbf{F}_i)^{-1}\mathbf{F}_i^T\mathbf{r}_{i,t}. \tag{7}$$

Each $\boldsymbol{\beta}_{i,t}$ is therefore a 3×1 vector, and concatenating $\boldsymbol{\beta}_{i,t}$ across t results in $\boldsymbol{\beta}_i$, a $3 \times T$ matrix, of which there are N. We then formed a tensor where each β_i is stacked, leading to a tensor with dimensions $3 \times T \times N$. For each of the 3 task variables, we found the time T where the norm of the regression coefficients, across all units, was largest. For the mth task variable, we denote the vector $\beta^m_{\max} \in \mathbb{R}^N$ to be a vector of coefficients that define a 1-dimensional projection of the neural population activity related to the mth task variable. These vectors are what we refer to as the task related axes. To orthogonalize these vectors, we performed QR decomposition on the stacked $\boldsymbol{\beta}_{\max}$ matrix $[\boldsymbol{\beta}^1_{\max}, \boldsymbol{\beta}^2_{\max}, \boldsymbol{\beta}^3_{\max}]$, which is an $N \times 3$ matrix. This decomposition finds orthogonal axes so that the axes would capture independent variance. The magnitude of the maximum regression coefficients $\boldsymbol{\beta}^m_{\max}$ are shown in Table 1, and are plotted in Fig. 3e. We verified that these axes were capturing a significant portion of total variance (Table 2) and that the axes were predominantly capturing corresponding task-related variability. For example, we verified that the direction axis predominantly captures direction related variability (Table 3).

For multi-area analyses, we separated the units for each area and found the task-relevant axes for this subset of units. For the inter-areal analyses, we found the color and direction axis using only the excitatory units because we were concerned the information contained in these units that would project to subsequent areas (only excitatory units had connections to the subsequent area). In all other analyses, all units were used to identify the axes.

We emphasize that the axes found by TDR explain behaviorally relevant or stimulus relevant variance. In our RNNs, overall variance includes components such as a general increase in firing rate, similar to a condition-independent signal found in PMd[76]. The TDR axes are task-relevant axes that reflect direction or signed color coherence, rather than axes that maximize captured variance. Consistent with this goal, the Area 1 color axis explains only 11% of overall variance but, 57% of variance associated with color discrimination, as shown in Tables 2 and 3. The Area 1 direction axis explains 68% of overall variance and 49% percent of the direction discrimination variance. Similarly, the Area 3 color axis explains 5.4% of overall variance and only 6.8% percent of the color discrimination variance while the direction axes explains 32% of overall variance and 77% of the direction discrimination variance (Tables 2, 3). The color axis explains little color variance in Area 3 because there is little color information in Area 3.

Axis overlap

To calculate axis overlap (Fig. 5g), we first note that the direction, color, and readout axes were unit length vectors. When vectors are unit length, their dot product is equal to the cosine of the angle between them, which is 1 for parallel vectors and 0 for orthogonal vectors. We therefore quantified axis overlap by taking the dot product between the direction (color) axis and the readout axis. We note that the readout axis only received projections from excitatory units, so we performed TDR to find the color and direction axis using only the excitatory units.

Analyses of inputs and activity

In order to disentangle the effects of external inputs and recurrence, in Fig. 4, Fig. 5d-f, and Fig. 8a, we evaluated the input contribution and overall activity. For Area 1, we defined the input contribution as $\mathbf{W}_{\text{in}}\mathbf{u}_t$, and for areas 2 and 3, we defined the input contribution as $\mathbf{W}_{12}\mathbf{r}_t^1$, and $\mathbf{W}_{23}\mathbf{r}_t^2$ respectively, where $\mathbf{r}_t^m$ denotes the activity of the units in area m. The activity $\mathbf{r}_t^m$ corresponds to the firing rate that experimentalists could measure, reflecting a combination of input and recurrent interactions. For constant inputs, a stable value of the activity implies there is little recurrent processing.

Null and Potent Analyses

To calculate the overlap between the color and direction axes with the null and potent spaces, we performed singular value decomposition on the inter-areal connections, $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$. $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$ were 80×80 matrices, and were full rank. Nevertheless, they had near some zero singular values, indicating that the effective rank of the matrix was less than 80. We defined the potent dimensions to be the top m right singular vectors, while the null dimensions were the remaining $80 - m$ right singular vectors.

We performed the analyses of Fig. 7a,b by varying the potent and null dimensions, sweeping m from 1 to 80. For each defined potent and null space, we calculated the axis overlap between the direction (or color) axis and the potent (or null) space by computing the L2-norm of the orthogonal projection (squared). We report the squared quantity because the expectation of the norm of a projection of a random vector onto an m-dimensional subspace of an n dimensional space is m/n. We include an approximation of the expectation of the projection of a random vector in Fig. 7a,b by averaging the projection of 100 random vectors. Our results show that the direction axis was always more aligned with potent dimensions than the color axis, irrespective of the choice of m, and that the direction axis was prefentially aligned with the top singular vector. These findings were consistent across 72 multi-area network instantiations.

Visualization of neural activity in a low dimensional space

The activity of multiple units on a single trial is high dimensional, with dimension equal to the number of units. To visualize the activity in a lower dimensional space, dimensionality reduction techniques can be used. In addition to TDR, we also utilized Principal Components Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (tSNE) to visualize neural activity in low-dimensional spaces.

PCA finds a linear low-dimensional projection of the high dimensional data that maximizes captured variance. We performed PCA on both the experimental data and RNN rates. PCA is an eigenvalue decomposition on the data covariance matrix. To calculate the covariance matrix of the data, we averaged responses across conditions. This reduces single trial variance and emphasizes variance across conditions. Firing rates were conditioned on reach direction and signed coherence. In both the experimental data and RNN rates, we had 28 conditions (14 signed coherences each for left and right reaches).

tSNE embeds high dimensional data in a low dimensional manifold that is nonlinear, enabling visualization of activity on a nonlinear manifold.. The tSNE embedding maintains relative distances between data points when reducing dimensionality, meaning that points closer in high dimensional space remain closer when viewed in a low dimensional manifold. We projected our data into a two dimensional manifold. We used the default parameters from the sickit-learn implementation (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html). The data visualized under tSNE was averaged in a window [-300ms, +100ms] around reaction time.

	color	direction	context
Area 1	11.4	7.4	5.8
Area 2	2.1	6.4	0.9
Area 3	1.2	9.1	0.6

Table 1: Magnitude of maximum beta coefficients

	color	direction	context
Area 1	10.4%	68.1%	3.1%
Area 2	15.4%	47.3%	1.3%
Area 3	5.4%	31.6%	8.2%

Table 2: Variance captured by axes across all trials

	Color		**Direction**	
	Color axis	Dir. axis	Color axis	Dir. axis
Area 1	56.7%	0.4%	21.3%	48.6%
Area 2	15.0%	11.7%	16.1%	54.2%
Area 3	6.8%	12.2%	0.5%	77.0%

Table 3: Variability of difference PSTHs captured

Supplementary information

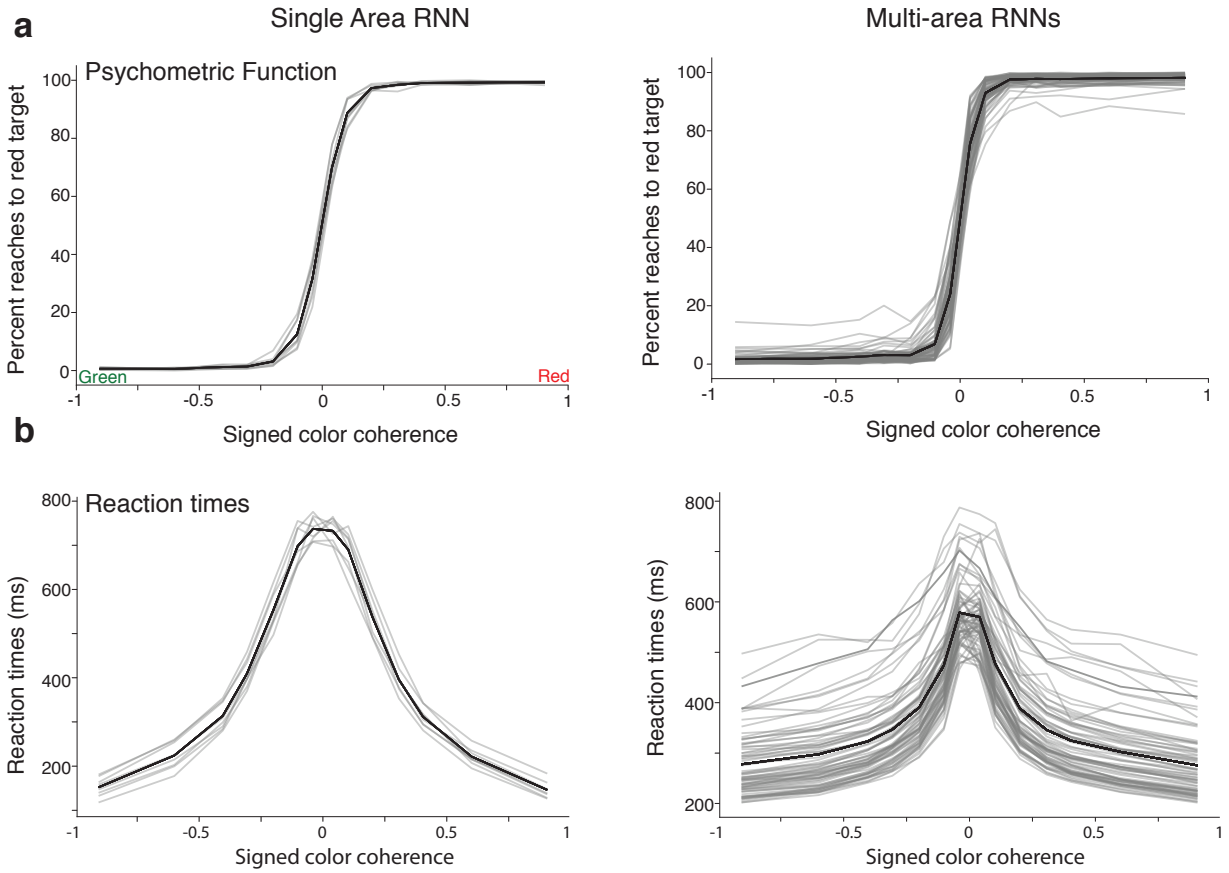

Figure S1: (**a**) Psychometric and (**b**) reaction time curves across single (left) and multi-area (right) RNNs trained for this study. Gray lines represent individual RNNs and the black solid line is the average across all RNNs.

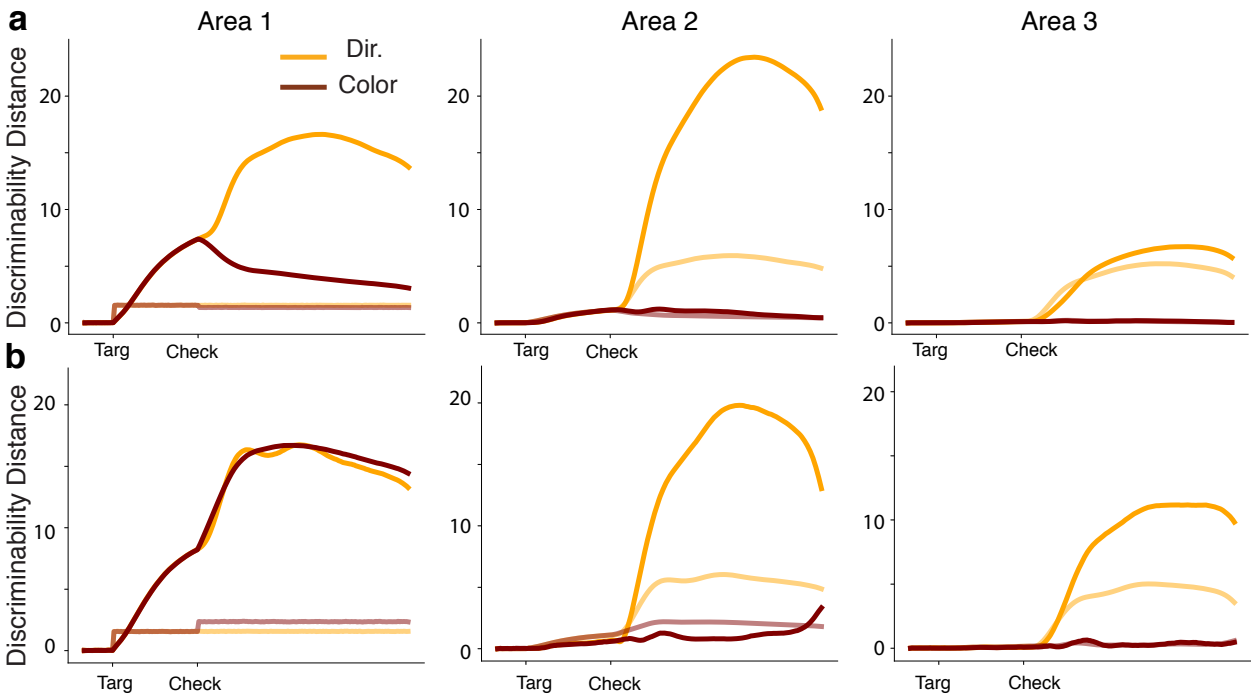

Figure S2: (**a**) The norm of the direction discriminability (left green - right green) and color discriminability (right green - right red) as a function of the processing area (same conventions as Fig. 4, but for the other discriminability). **(b)** If we consider the recurrent contribution $\mathbf{W}_{\text{rec}}\mathbf{r}_t$, as opposed to the activity $\mathbf{r}_t$, the structure of inter-areal attenuation and differential recurrent amplification is similar. In **(b)** we analyse the same discriminability as the text in Fig. 4.

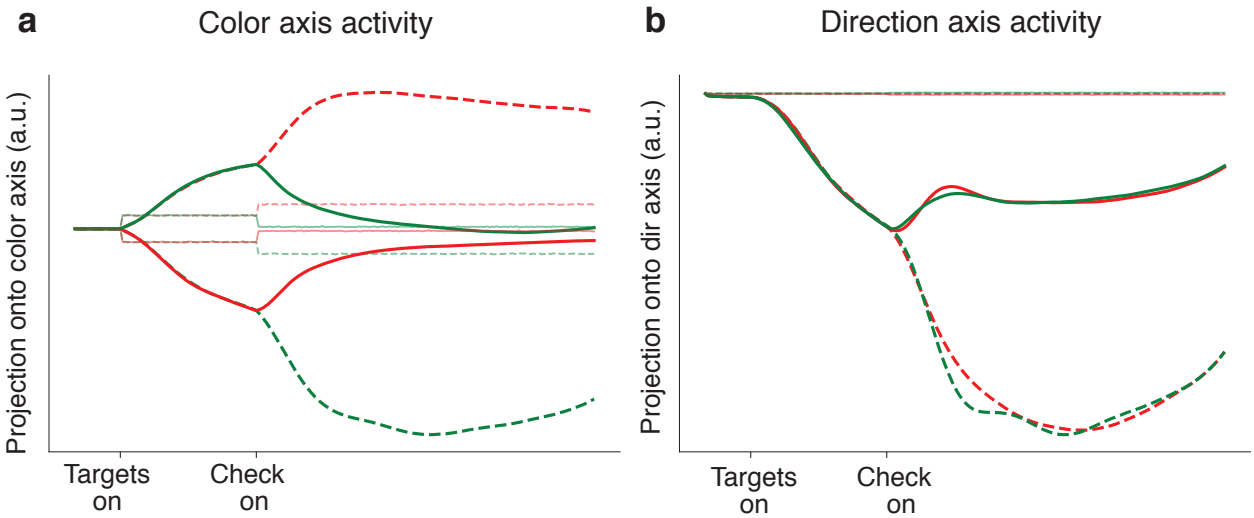

Figure S3: Finding the projections onto the color **(a)** and direction **(b)** axis in area 1 using only the excitatory units has little qualitative difference, in comparison to using all units (Fig. 5). Same conventions as Fig. 5.

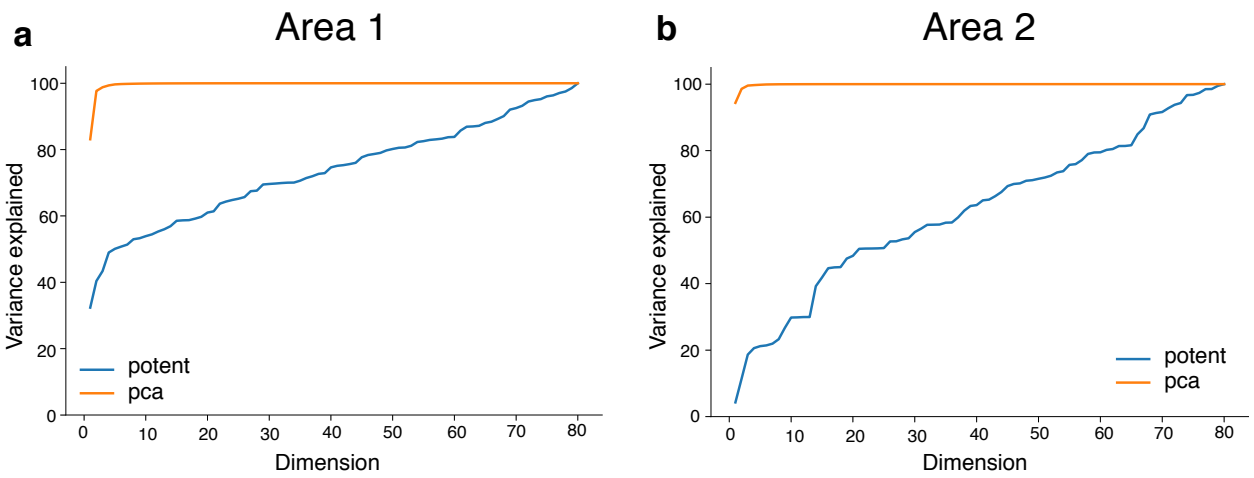

Figure S4: **(a)** Variance explained in Area 1 by the top principal components and top dimensions of potent space of $\mathbf{W}_{12}$, swept across all dimensions. **(b)** Variance explained in Area 2 by the top principal components and top dimensions of potent space pf $\mathbf{W}_{23}$, swept across all dimensions. These plots show that the connections between areas do not necessarily propagate the most dominant axes of variability in the source area to the downstream area.

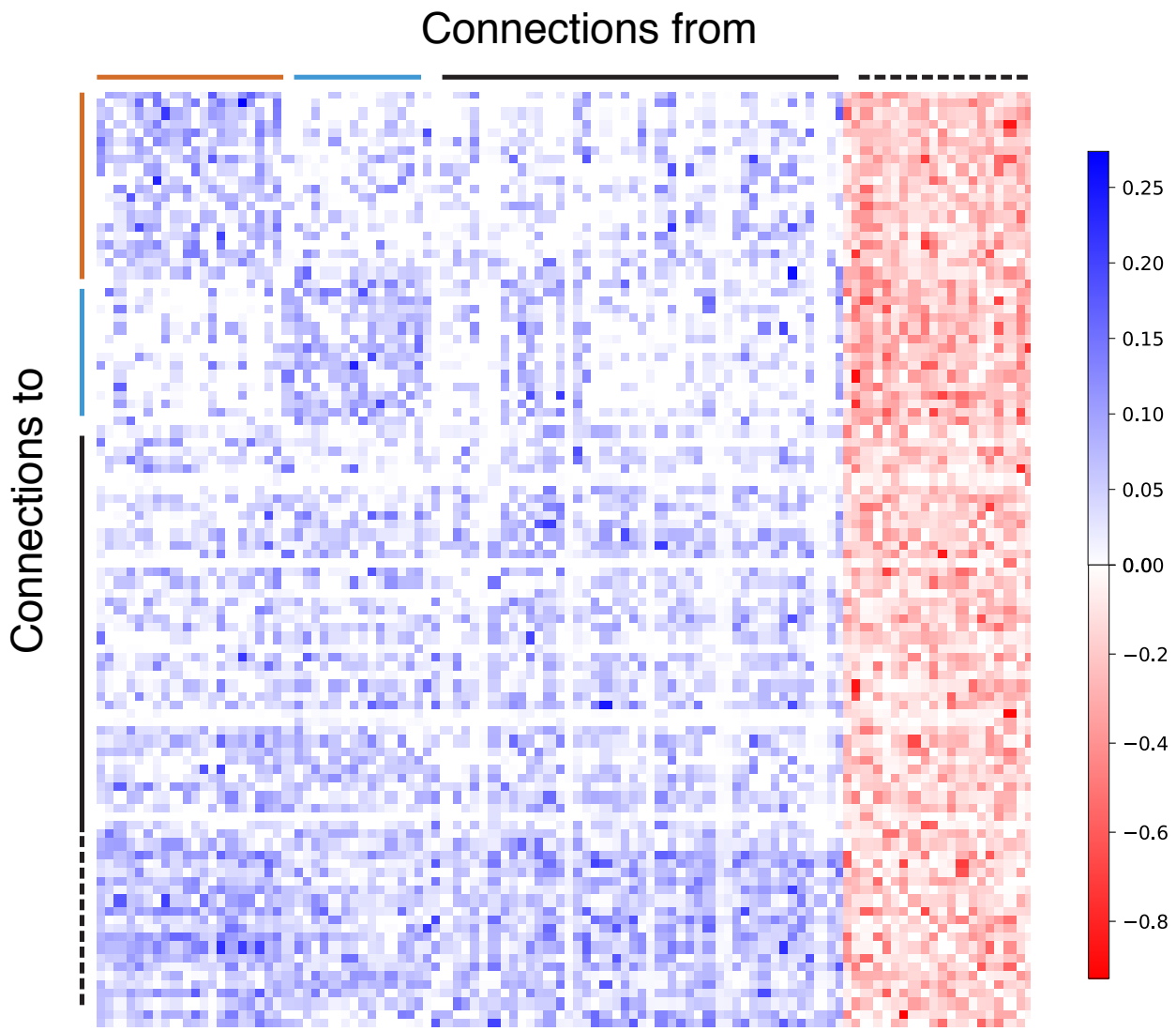

Figure S5: Full connectivity matrix of $\mathbf{W}_{\mathrm{rec}}$, reordered so that the structured excitatory components lie at the top left. The matrix is composed of a structured excitatory component (orange and blue), a set of random excitatory units (black), and a set of inhibitory units (dashed black), with little structure. The averaged connectivity matrix is shown in Fig. 8e. We note that since we did not see clear structure in the inhibitory pool (i.e mutual inhibtion), we therefore hypothesized common inhibtion.

Supplementary Note 1

Here we show that a nonlinearity is necessary to solve the task, proving that the task cannot be solved by the linear layer $\mathbf{W}_{\text{in}}$. Hence, we show that an "inputs only" hypothesis is not possible; there is a need for a nonlinearity.

First, we note that the Checkerboard task corresponds to an exclusive-or (XOR) problem. If we identify the two target configurations as 0 or 1 (corresponding to green on left, or green on right respectively, with the red target on the complement side), and the dominant checkerboard color as 0 or 1 (for green or red, respectively), then the output direction d (identified as 0: left, 1: right) can be seen be in Table S1.

If the representation $\mathbf{r}$ was purely input driven, then:

$$\mathbf{r} = \mathbf{W}_{\text{in}}\mathbf{u}, \tag{8}$$

Our readout was a linear readout of the rates, i.e:

$$d = \mathbf{W}_{\text{out}}\mathbf{r} \tag{9}$$

The inputs $\mathbf{u}$ are the four dimensional input we trained with. But $\mathbf{u}$ is a linear transformation of two variables: the target orientation θ, and checkerboard color c, which each can take two values. That is, if we let $\mathbf{q} = [\theta, c]$, then, the inputs could be written as a linear transformation of $\mathbf{q}$:

$$\mathbf{u} = \mathbf{W}\mathbf{q}, \tag{10}$$

where $\mathbf{W}$ is a linear transformation. Since the mappings from $\mathbf{q}$ to d are all linear, they can be combined into a single linear transformation $\tilde{\mathbf{W}}$, i.e.,

$$d = \mathbf{W}_{\text{out}}\mathbf{W}_{\text{in}}\mathbf{W}\mathbf{q} = \tilde{\mathbf{W}}\mathbf{q}. \tag{11}$$

It is not possible for a linear classifier to solve the XOR problem by classifying correct outputs[74]. Hence, the trained RNNs cannot purely be input driven, and requires nonlinearity from the recurrent interactions to solve the task.

target configuration	color	direction
0	0	0
0	1	1
1	0	1
1	1	0

Table S1: Checkerboard task truth table

Word Counts

- Abstract: 198 words
- Introduction: 795 words
- Results: 4741 words
- Methods: 3155 words
- Discussion: 1792 words

Acknowledgments

MK was supported by the National Sciences and Engineering Research Council (NSERC). CC was supported by a NIH/NINDS R00 award R00NS092972 and by the Moorman-Simon Interdisciplinary Career Development Professorship from Boston University, and the NARSAD Young Investigator Award from the Brain and Behavior Research Foundation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We thank Laura Driscoll for helpful comments on the manuscript as well as Krishna V. Shenoy and William T. Newsome for helpful discussions.

Author contributions

JCK and CC conceived of the study. MK and JCK trained RNNs and analyzed networks. MK performed the multi-area computation analyses. CC collected experimental data in PMd. All authors came up with analyses and conceptual frameworks to understand computation in the RNNs, discussed results, and wrote and edited the manuscript.

References

1. Szczepanski, SM, Konen, CS, and Kastner, S (2010). *Mechanisms of Spatial Attention Control in Frontal and Parietal Cortex*. Journal of Neuroscience, 30(1):148–160. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3862-09.2010. URL https://www.jneurosci.org/content/30/1/148.
2. Kalaska, J and Crammond, D (1992). *Cerebral cortical mechanisms of reaching movements*. Science, 255(5051):1517–1523. ISSN 0036-8075. doi:10.1126/science.1549781. URL https://science.sciencemag.org/content/255/5051/1517.
3. Remington, ED, Narain, D, Hosseini, EA, and Jazayeri, M (2018). *Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics*. Neuron, 98(5):1005–1019.e5.
4. Yamagata, T, Nakayama, Y, Tanji, J, and Hoshi, E (2012). *Distinct information representation and processing for goal-directed behavior in the dorsolateral and ventrolateral prefrontal cortex and the dorsal premotor cortex*. Journal of Neuroscience, 32(37):12934–12949.
5. Yau, JM, DeAngelis, GC, and Angelaki, DE (2015). *Dissecting neural circuits for multisensory integration and crossmodal processing*. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1677):20140203. doi:10.1098/rstb.2014.0203. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2014.0203.
6. Cisek, P (2012). *Making decisions through a distributed consensus*. Current opinion in neurobiology, 22(6):927–936.
7. Pinto, L, Rajan, K, DePasquale, B, Thiberge, SY, Tank, DW, and Brody, CD (2019). *Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions*. Neuron. doi:10.1016/j.neuron.2019.08.025. URL https://doi.org/10.1016/j.neuron.2019.08.025.
8. Siegel, M, Buschman, TJ, and Miller, EK (2015). *Cortical information flow during flexible sensorimotor decisions*. Science, 348(6241):1352–1355.
9. Freedman, DJ and Assad, JA (2016). *Neuronal mechanisms of visual categorization: an abstract view on decision making*. Annual review of neuroscience, 39:129–147.
10. Brown, EN, Kass, RE, and Mitra, PP (2004). *Multiple neural spike train data analysis: state-of-the-art and future challenges*. Nature neuroscience, 7(5):456.
11. Gerstein, GL and Perkel, DH (1969). *Simultaneously recorded trains of action potentials: analysis and functional interpretation*. Science, 164(3881):828–830.
12. Ghazanfar, AA, Chandrasekaran, C, and Logothetis, NK (2008). *Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys*. Journal of Neuroscience, 28(17):4457–4469.
13. Buschman, TJ and Miller, EK (2007). *Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices*. science, 315(5820):1860–1862.
14. Saalmann, YB, Pigarev, IN, and Vidyasagar, TR (2007). *Neural mechanisms of visual attention: how top-down feedback highlights relevant locations*. Science, 316(5831):1612–1615.
15. Gregoriou, GG, Gotts, SJ, Zhou, H, and Desimone, R (2009). *High-frequency, long-range coupling between prefrontal and visual cortex during attention*. science, 324(5931):1207–1210.
16. Semedo, JD, Zandvakili, A, Machens, CK, Yu, BM, and Kohn, A (2019). *Cortical Areas Interact through a Communication Subspace*. Neuron, 102(1):249–259.e4. ISSN 0896-6273. doi:10.1016/j.neuron.2019.01.026. URL https://doi.org/10.1016/j.neuron.2019.01.026.
17. Tauste Campo, A, Martinez-Garcia, M, Nácher, V, Luna, R, Romo, R, and Deco, G (2015). *Task-driven intra- and interarea communications in primate cerebral cortex*. Proceedings of the National Academy of Sciences, 112(15):4761–4766. ISSN 0027-8424. doi:10.1073/pnas.1503937112. URL https://www.pnas.org/content/112/15/4761.
18. Yamins, DLK, Hong, H, Cadieu, CF, Solomon, EA, Seibert, D, and DiCarlo, JJ (2014). *Performance-*

optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A., 111(23):8619–8624.

19. Yamins, DLK and DiCarlo, JJ (2016). *Using goal-driven deep learning models to understand sensory cortex*. Nat. Neurosci., 19(3):356–365.
20. Mante, V, Sussillo, D, Shenoy, KV, and Newsome, WT (2013). *Context-dependent computation by recurrent dynamics in prefrontal cortex*. Nature, 503(7474):78–84.
21. Song, HF, Yang, GR, and Wang, XJ (2016). *Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework*. PLoS Comput. Biol., 12(2):1–30.
22. Rajan, K, Harvey, CD, and Tank, DW (2016). *Recurrent Network Models of Sequence Generation and Memory*. Neuron, 90(1):128–142.
23. Chaisangmongkon, W, Swaminathan, SK, Freedman, DJ, and Wang, XJ (2017). *Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions*. Neuron, 93(6):1504–1517.e4.
24. Song, HF, Yang, GR, and Wang, XJ (2017). *Reward-based training of recurrent neural networks for cognitive and value-based tasks*. Elife, 6.
25. Yang, GR, Joglekar, MR, Francis Song, H, Newsome, WT, and Wang, XJ (2019). *Task representations in neural networks trained to perform many cognitive tasks*.
26. Orhan, AE and Ma, WJ (2019). *A diverse range of factors affect the nature of neural representations underlying short-term memory*. Nature Neuroscience, 22(2):275–283.
27. Laje, R and Buonomano, DV (2013). *Robust timing and motor patterns by taming chaos in recurrent neural networks*. Nature neuroscience, 16(7):925.
28. Goudar, V and Buonomano, DV (2018). *Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks*. eLife, 7:e31134. ISSN 2050-084X. doi:10.7554/eLife.31134. URL https://doi.org/10.7554/eLife.31134.
29. Banino, A, Barry, C, Uria, B, Blundell, C, Lillicrap, T, Mirowski, P, Pritzel, A, Chadwick, MJ, Degris, T, Modayil, J, Wayne, G, Soyer, H, Viola, F, Zhang, B, Goroshin, R, Rabinowitz, N, Pascanu, R, Beattie, C, Petersen, S, Sadik, A, Gaffney, S, King, H, Kavukcuoglu, K, Hassabis, D, Hadsell, R, and Kumaran, D (2018). *Vector-based navigation using grid-like representations in artificial agents*. Nature.
30. Hennequin, G, Vogels, TP, and Gerstner, W (2014). *Optimal control of transient dynamics in balanced networks supports generation of complex movements*. Neuron, 82(6):1394–1406.
31. Sussillo, D, Churchland, MM, Kaufman, MT, and Shenoy, KV (2015). *A neural network that finds a naturalistic solution for the production of muscle activity*. Nat. Neurosci., 18(7):1025–1033.
32. Michaels, JA, Dann, B, and Scherberger, H (2016). *Neural population dynamics during reaching are better explained by a dynamical system than representational tuning*. PLoS Comput. Biol., 12(11):e1005175.
33. Stroud, JP, Porter, MA, Hennequin, G, and Vogels, TP (2018). *Motor primitives in space and time via targeted gain modulation in cortical networks*. Nature neuroscience, 21(12):1774.
34. Barak, O (2017). *Recurrent neural networks as versatile tools of neuroscience research*. Current Opinion in Neurobiology, 46:1 – 6. ISSN 0959-4388. doi:https://doi.org/10.1016/j.conb.2017.06.003. URL http://www.sciencedirect.com/science/article/pii/S0959438817300429. Computational Neuroscience.
35. Mastrogiuseppe, F and Ostojic, S (2018). *Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks*. Neuron.
36. Sussillo, D and Barak, O (2013). *Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks*. Neural Comput., 25(3):626–649.
37. Chandrasekaran, C, Peixoto, D, Newsome, WT, and Shenoy, KV (2017). *Laminar differences in decision-related neural activity in dorsal premotor cortex*. Nat. Commun., 8(1):614.

38. Kaufman, MT, Churchland, MM, Ryu, SI, and Shenoy, KV (2015). *Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex*. Elife, 4:1–21.
39. Coallier, É, Michelet, T, and Kalaska, JF (2015). *Dorsal premotor cortex: neural correlates of reach target decisions based on a color-location matching rule and conflicting sensory evidence*. American Journal of Physiology-Heart and Circulatory Physiology.
40. Gold, JI and Shadlen, MN (2007). *The neural basis of decision making*. Annual review of neuroscience, 30.
41. Romo, R and de Lafuente, V (2013). *Conversion of sensory signals into perceptual decisions*. Progress in Neurobiology, 103:41 – 75. ISSN 0301-0082. doi:https://doi.org/10.1016/j.pneurobio.2012.03.007. URL http://www.sciencedirect.com/science/article/pii/S0301008212000391. Conversion of Sensory Signals into Perceptions, Memories and Decisions.
42. Kaufman, MT, Churchland, MM, Ryu, SI, and Shenoy, KV (2014). *Cortical activity in the null space: permitting preparation without movement*. Nat. Neurosci., 17(3):440–448.
43. Miri, A, Warriner, CL, Seely, JS, Elsayed, GF, Cunningham, JP, Churchland, MM, and Jessell, TM (2017). *Behaviorally selective engagement of short-latency effector pathways by motor cortex*. Neuron, 95(3):683–696.
44. Wang, M, Montanède, C, Chandrasekaran, C, Peixoto, D, Shenoy, KV, and Kalaska, JF (2019). *Macaque dorsal premotor cortex exhibits decision-related activity only when specific stimulus–response associations are known*. Nature Communications, 10(1):1793. ISSN 2041-1723. doi:10.1038/s41467-019-09460-y. URL https://doi.org/10.1038/s41467-019-09460-y.
45. Kingma, DP and Ba, J (2014). *Adam: A Method for Stochastic Optimization*. arXiv preprint arXiv:1412.6980.
46. Pascanu, R, Mikolov, T, and Bengio, Y (2013). *On the difficulty of training recurrent neural networks*. In *International Conference on Machine Learning*, pages 1310–1318.
47. Wang, M, Montanede, C, Chandrasekaran, C, Peixoto, D, Shenoy, KV, and Kalaska, JF (2019). *Macaque dorsal premotor cortex exhibits decision-related activity only when specific stimulus-response associations are known*. Nature Communications, pages –in Press.
48. Wise, SP and Murray, EA (2000). *Arbitrary associations between antecedents and actions*. Trends in Neurosciences, 23(6):271 – 276. ISSN 0166-2236. doi:https://doi.org/10.1016/S0166-2236(00)01570-8. URL http://www.sciencedirect.com/science/article/pii/S0166223600015708.
49. Hoshi, E (2013). *Cortico-basal ganglia networks subserving goal-directed behavior mediated by conditional visuo-goal association*. Front. Neural Circuits, 7(October):158.
50. Badre, D, Hoffman, J, Cooney, JW, and D'esposito, M (2009). *Hierarchical cognitive control deficits following damage to the human frontal lobe*. Nature neuroscience, 12(4):515.
51. Barbas, H and Pandya, DN (1987). *Architecture and frontal cortical connections of the premotor cortex (area 6) in the rhesus monkey*. Journal of Comparative Neurology, 256(2):211–228. doi:10.1002/cne.902560203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.902560203.
52. Luppino, G, Rozzi, S, Calzavara, R, and Matelli, M (2003). *Prefrontal and agranular cingulate projections to the dorsal premotor areas F2 and F7 in the macaque monkey*. European Journal of Neuroscience, 17(3):559–578. doi:10.1046/j.1460-9568.2003.02476.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1460-9568.2003.02476.x.
53. Markov, NT, Ercsey-Ravasz, MM, Ribeiro Gomes, AR, Lamy, C, Magrou, L, Vezoli, J, Misery, P, Falchier, A, Quilodran, R, Gariel, MA, Sallet, J, Gamanut, R, Huissoud, C, Clavagnier, S, Giroud, P, Sappey-Marinier, D, Barone, P, Dehay, C, Toroczkai, Z, Knoblauch, K, Van Essen, DC, and Kennedy, H (2014). *A weighted and directed interareal connectivity matrix for macaque cerebral cortex*. Cereb. Cortex, 24(1):17–36.
54. Van der Maaten, L and Hinton, G (2008). *Visualizing Data using t-SNE*. Journal of Machine Learning . . . , 9:2579–2605.

55. Svoboda, K and Li, N (2018). *Neural mechanisms of movement planning: motor cortex and beyond*. Current Opinion in Neurobiology, 49:33 – 41. ISSN 0959-4388. doi:https://doi.org/10.1016/j.conb.2017.10.023. URL http://www.sciencedirect.com/science/article/pii/S0959438817302283. Neurobiology of Behavior.
56. Cunningham, JP and Yu, BM (2014). *Dimensionality reduction for large-scale neural recordings*. Nat. Neurosci., 17(11):1500–1509.
57. Gallego, JA, Perich, MG, Miller, LE, and Solla, SA (2017). *Neural Manifolds for the Control of Movement*. Neuron, 94(5):978–984.
58. Pandarinath, C, Ames, KC, Russo, AA, Farshchian, A, Miller, LE, Dyer, EL, and Kao, JC (2018). *Latent Factors and Dynamics in Motor Cortex and Their Application to Brain-Machine Interfaces*. J. Neurosci., 38(44):9390–9401.
59. Afshar, A, Santhanam, G, Yu, BM, Ryu, SI, Sahani, M, and Shenoy, KV (2011). *Single-trial neural correlates of arm movement preparation*. Neuron, 71(3):555–564.
60. Churchland, MM, Cunningham, JP, Kaufman, MT, Foster, JD, Nuyujukian, P, Ryu, SI, and Shenoy, KV (2012). *Neural population dynamics during reaching*. Nature, 487(7405):51–56.
61. Ames, KC, Ryu, SI, and Shenoy, KV (2014). *Neural dynamics of reaching following incorrect or absent motor preparation*. Neuron, 81(2):438–451.
62. Elsayed, GF, Lara, AH, Kaufman, MT, Churchland, MM, and Cunningham, JP (2016). *Reorganization between preparatory and movement population responses in motor cortex*. Nat. Commun., 7:13239.
63. Gallego, JA, Perich, MG, Naufel, SN, Ethier, C, Solla, SA, and Miller, LE (2018). *Cortical population activity within a preserved neural manifold underlies multiple motor behaviors*. Nat. Commun., 9(1):4233.
64. Li, N, Daie, K, Svoboda, K, and Druckmann, S (2016). *Robust neuronal dynamics in premotor cortex during motor planning*. Nature, 532(7600):459–464.
65. Harvey, CD, Coen, P, and Tank, DW (2012). *Choice-specific sequences in parietal cortex during a virtual-navigation decision task*. Nature, 484:62–68.
66. Marcos, AS and Harvey, CD (2016). *History-dependent variability in population dynamics during evidence accumulation in cortex*. Nat. Neurosci., 19(12):1672–1680.
67. Michaels, JA, Schaffelhofer, S, Agudelo-Toro, A, and Scherberger, H (2019). *A neural network model of flexible grasp movement generation*. bioRxiv, page 742189.
68. Cisek, P and Kalaska, JF (2005). *Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action*. Neuron, 45(5):801–814.
69. Shwartz-Ziv, R and Tishby, N (2017). *Opening the Black Box of Deep Neural Networks via Information*. CoRR, abs/1703.00810. URL http://arxiv.org/abs/1703.00810.
70. Achille, A and Soatto, S (2018). *Emergence of Invariance and Disentanglement in Deep Representations*. Journal of Machine Learning Research, 19(50):1–34. URL http://jmlr.org/papers/v19/17-646.html.
71. Tishby, N, Pereira, FC, and Bialek, W (1999). *The Information Bottleneck Method*. pages 368–377. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.9882.
72. Stavisky, SD, Kao, JC, Ryu, SI, and Shenoy, KV (2017). *Motor Cortical Visuomotor Feedback Activity Is Initially Isolated from Downstream Targets in Output-Null Neural State Space Dimensions*. Neuron, 95(1):195–208.e9.
73. Erlich, JC, Brunton, BW, Duan, CA, Hanks, TD, and Brody, CD (2015). *Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat*. Elife, 4:e05457.
74. Goodfellow, I, Bengio, Y, and Courville, A (2016). *Deep Learning*. MIT Press.
75. Markov, NT, Vezoli, J, Chameau, P, Falchier, A, Quilodran, R, Huissoud, C, Lamy, C, Misery, P, Giroud, P, Ullman, S, Barone, P, Dehay, C, Knoblauch, K, and Kennedy, H (2014). *Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex*. Journal of Comparative

Neurology, 522(1):225–259. doi:10.1002/cne.23458. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.23458.

76. Kaufman, MT, Seely, JS, Sussillo, D, Ryu, SI, Shenoy, KV, and Churchland, MM (2016). *The largest response component in motor cortex reflects movement timing but not movement type*. eNeuro, 3(August):ENEURO.0085–16.2016.