A mechanistic multi-area recurrent network model of decision-making

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recurrent neural networks (RNNs) trained on neuroscience-based tasks have been widely used as models for cortical areas performing analogous tasks. However, very few tasks involve a single cortical area, and instead require the co-ordination of multiple brain areas. Despite the importance of multi-area computation, there is a limited understanding of the principles underlying such computation. We propose to use multi-area RNNs with neuroscience-inspired architecture constraints to derive key features of multi-area computation. In particular, we show that incorporating multiple areas and Dale's Law is critical for biasing the networks to learn biologically plausible solutions. Additionally, we leverage the full observability of the RNNs to show that output-relevant information is preferentially propagated between areas. These results suggest that cortex uses modular computation to generate minimal sufficient representations of task information. More broadly, our results suggest that constrained multi-area RNNs can produce experimentally testable hypotheses for computations that occur within and across multiple brain areas, enabling new insights into distributed computation in neural systems.

1 Introduction

Decision-making, multisensory integration, attention, motor control, and timing emerge from the coordination of multiple interconnected brain areas [1–9]. While neural activity 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]. Understanding the neural bases of these behaviors necessitates an understanding of the intra- and inter-area dynamics, that is, how neural activity evolves within and between brain areas. However, we currently lack clear hypotheses for how distinct brain-area dynamics and connectivity relate to computation. To address this gap, we use multi-area recurrent neural networks (RNNs) to model, probe, and gain insight into how behaviorally relevant computations emerge within and across multiple brain areas.

Optimized feedforward and recurrent neural networks have been used for machine learning but are also emerging tools to model computations associated with visual [10, 11], cognitive [12], timing [3, 13], navigation [14], and motor tasks [15, 16]. RNNs transform experimenter-designed task inputs into behavior-related outputs through recurrent dynamics. Its artificial units often exhibit heterogeneous responses and population dynamics observed in brain areas implicated in cognitive and motor tasks [12, 15, 17–20]. If artificial units resemble cortical neurons, RNNs are subsequently analyzed to propose hypotheses for how a local computation occurs in a brain area [3, 12, 18, 21]. In comparison to direct experimental recordings, an important advantage of RNNs is that the activity of all artificial units and their recurrent connectivity are fully observed. It is therefore possible to engineer [22] and reverse engineer [23] RNNs by analyzing their activity and recurrent connectivity, providing mechanistic insight into cortical computation [3, 12, 21].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

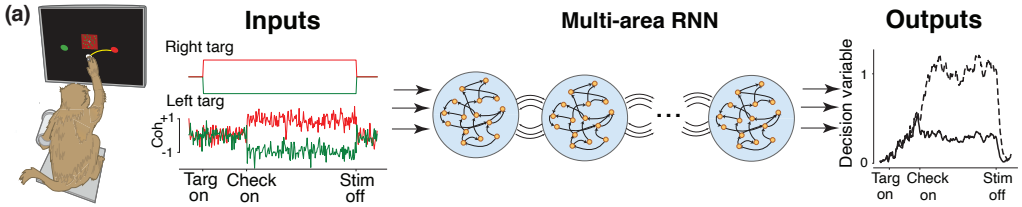

Figure 1: **Task. (a)** RNN configuration. The RNN receives 4 inputs. The first two inputs indicate the identity of the left and right targets, which can be red or green. These inputs are noiseless. The last two inputs indicate the value of the signed color coherence (proportional to amount of red in checkerboard) and negative signed color coherence (proportional to amount of green in checkerboard). We added 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.

Traditionally, RNNs have provided insight into local computations, and there has been limited insight into multi-area computation [7, 19]. To study multi-area computation, we explicitly constrained RNNs to have multiple recurrent areas, which we refer to as multi-area RNNs. We used these multi-area RNNs to study decision-making, a cognitive process known to involve multiple areas including the prefrontal, parietal, and premotor cortex [7, 9, 12, 24–28]. Multi-area RNNs enable us to investigate several questions. Most broadly, what are the roles of within-area dynamics and inter-area connections in mediating distributed computations? How does the dimensionality and dynamics of neural computation differ across areas? What role do inter-area feedforward and feedback connections play in propagating information and rejecting noise? How do intra-area dynamics and inter-area connections coordinate to solve a task?

We use multi-area RNNs to study these questions in a decision-making task where premotor cortex and upstream areas are known to perform distinct computations. We trained multi-area RNNs to perform a perceptual decision-making task (Checkerboard Task) and compared their activity to monkey neuron recordings from the dorsal premotor cortex (PMd). We found that, when incorporating Dale's law and anatomically-informed levels of feedforward inhibition into training, PMd-resembling dynamics emerged in multi-area RNNs. Specifically, the multi-area RNN's output area (1) resembled PMd in single unit statistics and neural population activity, and (2) only retained the "output relevant" signals. Inter-area connections preferentially propagated these output relevant signals while attenuating output irrelevant signals. Our models and analyses provide a framework for studying distributed computations involving multiple areas in neural systems.

2 Motivation: Decision-making involves multiple brain areas

2.1 Checkerboard Task

In the "Checkerboard Task" [24, 26], shown in Fig. 1, a monkey was first shown left and right targets whose color (red and green) was random on each trial. The monkey 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 colors were random on each trial, this task separates the reach direction decision from the color decision [29]. This task enables studying how information related to the selection of the color of the target and information related to the direction of the reach is represented.

2.2 PMd Data during Checkerboard Task

We analyzed the activity of neurons from the dorsal premotor cortex (PMd) in monkeys performing the Checkerboard Task [30]. In the Checkerboard Task, neural activity in PMd, an area associated with somatomotor decisions, principally reflects the direction decision (left or right) and has minimal representations associated with the dominant color of the checkerboard (red or green) [26, 30, 31]. To

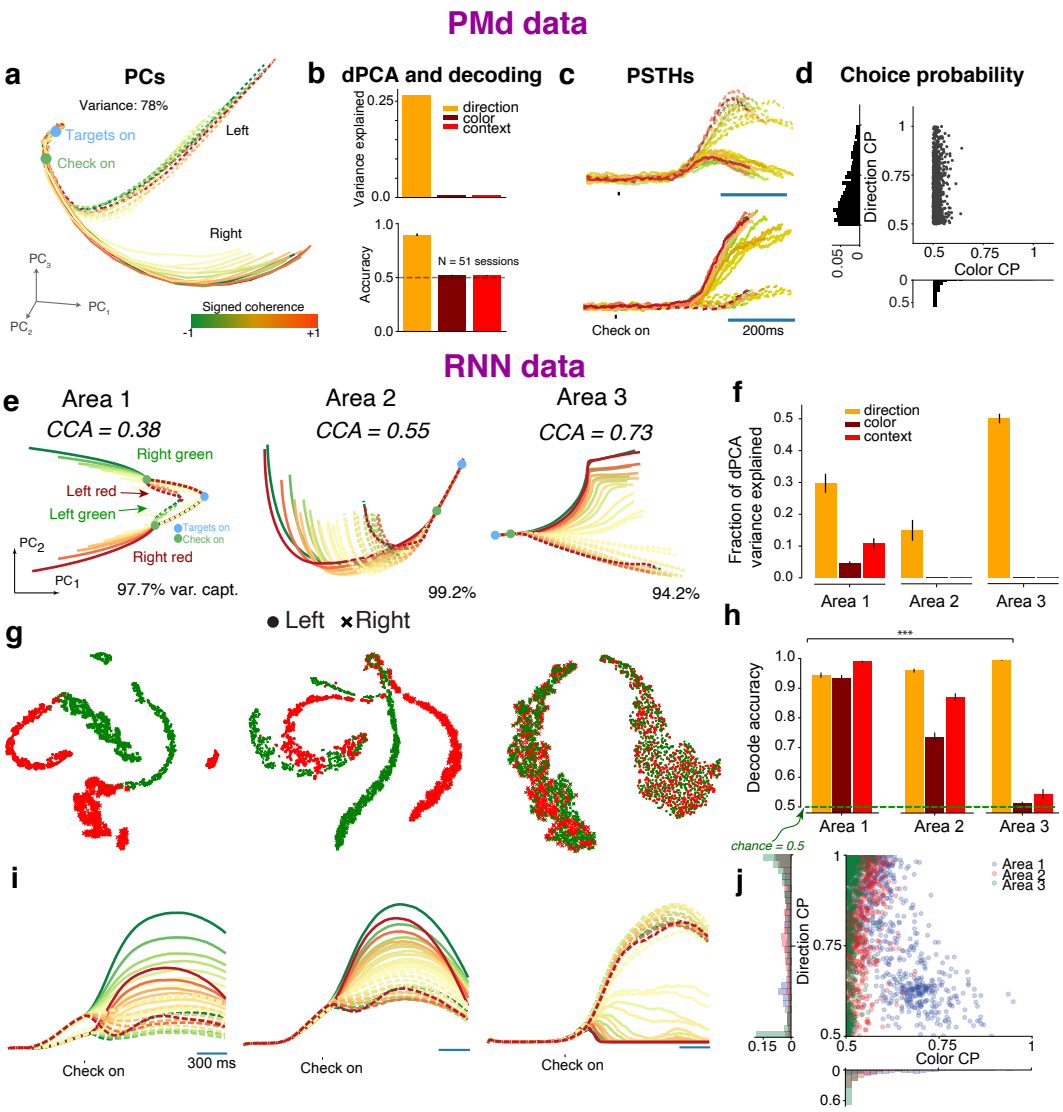

Figure 2: **PMd-resembling dynamics emerge in neuroscience constrained RNNs.** **(a)** PMd neural trajectories in the top 3 PCs. Color reflects signed color coherence, with darker shades of red (green) indicating more red (green) checkerboards. Right (left) reaches are denoted by solid (dotted) lines. **(b)** (Top) Variance captured by dPCA axes for the color decision, target configuration (context), and direction decision. (Bottom) Decode accuracy of the direction decision, color decision, and context in PMd sessions with U-probes and multiple neurons. **(c)** Representative PMd PSTHs aligned to checkerboard onset. **(d)** Direction and color choice probability (CP) for all recorded PMd units. **(e)** Neural trajectories in the top 2 principal components for each RNN area. **(f)** Variance captured by dPCA axes for color, context, and direction. Area 1 contains color, direction and context information, whereas Area 3 contains minimal variance for color and context. Note dPCA is a linear method that depends on marginalization. **(g)** Non-linear tSNE embedding of peri-movement activity in each area. Each dot is a trial, with red or green denoting the color decision and '.' or 'x' denoting the direction decision. Unlike Areas 1 and 2, Area 3 only had two clusters separated based on the direction decision. **(h)** Decode accuracy of direction, color, and context in all three areas. **(i)** Example PSTHs in each area. **(j)** Choice probabilities for units in all areas (pooled over 8 RNNs).

summarize this phenomenon, we show the principal components (PCs) of the PMd neural population activity in Fig. 2a. These PC trajectories separated based on the eventual reach direction (right reaches in solid, left in dotted), not color (red and green overlapping). We identified principal axes via demixed

PCA (dPCA [32]) that maximized variance related to the target configuration (context), color decision, and direction decision (see Methods). The direction axes captured significant variance (26.7%) while the color and context axes captured minimal variance (0.7%, 0.5%, respectively), as shown in Fig. 2e. It is possible, however, that there is direction-dependent color variance that is averaged away during marginalization when computing the dPCA variance [32]. Given simultaneously recorded data, a more appropriate measure of representation is the decode accuracy of direction, color and context. Across sessions where we analyzed multiple simultaneously recorded units from U-probes, the direction decision could be decoded from PMd activity significantly above chance (accuracy: 0.89, $p < 0.01$, bootstrap), but the color decision and context decode accuracy were not significantly above chance in any session (overall accuracies: 0.52 and 0.52, respectively, Fig. 2b, bottom).

Single neurons also had minimal color separation in individual PSTHs (e.g., Fig. 2c). To summarize this effect in single neurons, we computed the choice probabilities (CPs) reflecting how well the direction decision (direction CP) and color decision (color CP) could be decoded. PMd units generally had near chance color CP (0.5), but moderate to high direction CP, as shown in Fig. 2d. Together, these results demonstrate that PMd largely represents direction-related signals, but not the color decision or target configuration context. Since PMd activity minimally represents the color of the checkerboard or the target configuration, we reasoned that checkerboard and target inputs are transformed into a direction signal upstream of PMd and that multiple brain areas are necessary for solving this task. Brain areas, including the dorsolateral prefrontal cortex (DLPFC), and the ventrolateral prefrontal cortex (VLPFC), have been implicated in related sensorimotor transformations [4, 9, 33–35].

3 Results

Because the Checkerboard Task involves multiple brain areas, we reasoned that a single-area RNN would not resemble PMd recordings. We trained traditional single-area RNNs to perform the Checkerboard Task. The RNN processed the target context and checkerboard coherence to produce two decision variables corresponding to accumulated evidence for a left and right decision (Fig. 1). All training details are discussed in Appendix 5.4. We found that these RNN representations mixed color and direction information, as summarized in Appendix Fig 8, and therefore did not resemble PMd activity. This led us to study multi-area RNN models performing the Checkerboard Task, which turn out to accurately model PMd activity.

3.1 PMd-like representations emerge in optimized multi-area RNNs with neuroscience constraints

Given the anatomical and physiological evidence suggesting that multiple brain areas are implicated in the CB task, we hypothesized that the last area of an optimized multi-area RNN would more closely resemble PMd, receiving transformed direction signals computed using the checkerboard coherence and target configuration from upstream areas. We trained multi-area RNNs to perform the Checkerboard Task. The RNN had 3 areas, obeyed Dale's law [36], and had approximately 10% feedforward and 5% feedback connections between areas based on projections between prefrontal and premotor cortex in a macaque atlas [37]. All training details, including input and output configuration, optimization, and architectural constraints, are described in the Appendix 5.4.

The 3-area RNN had qualitatively different population trajectories across areas, shown in Fig. 2e. Area 1 had four distinct trajectory motifs corresponding to the four possible task outcomes (combinations of left vs right and red vs green decisions). PC_1 primarily varied with direction, while PC_2 varied with both the target context and red versus green checkerboards. In contrast, Area 2 and Area 3 population trajectories primarily separated on direction, not color, like in PMd. Area 3 trajectories most strongly resembled PMd trajectories (canonical correlations, $r = 0.38, 0.55, 0.73$ for Areas 1, 2, and 3; see Appendix).

We quantified the variance captured by dPCA principal axes for the context, color, and direction axis. We found that color axis variance decreased in later areas (Area 1: 5.6%, Area 2: 0.13%, Area 3: 0.07%, Fig. 2f). In contrast, Area 3 had the largest direction axis variance (Area 1: 30.9%, Area 2: 18.2%, Area 3: 48.5%, Fig. 2f). An important assumption of dPCA is that the neural activity can be decomposed as a sum of terms that depend solely on particular task variables [38]. The color variance found by dPCA indicate that color, on its own, did not account for a large fraction of the overall

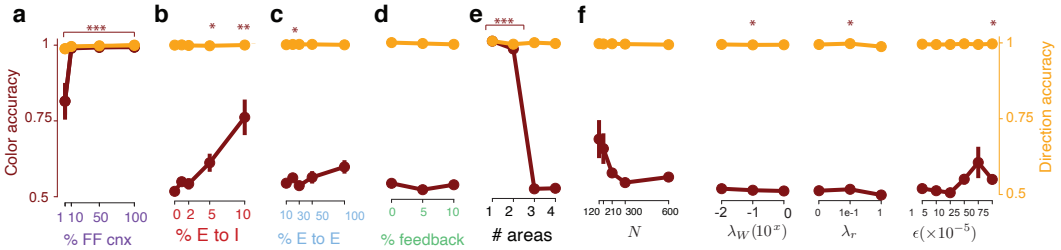

Figure 3: **PMd-resembling dynamics emerge in neuroscience constrained RNNs. (a)** We trained 3-area RNNs without explicit excitatory (E) or inhibitory (I) neurons. Inputs projected onto Area 1, and outputs were read out from Area 3. We varied the percentage of feedforward connections and computed the color and direction accuracy in Area 3. At 1% feedforward connections, color could still be significantly decoded above chance. Dots are the mean across networks and error bars are s.e.m. For significance, * is $p < 0.05$, ** is $p < 0.01$, and *** is $p < 0.001$ (with appropriate correction for multiple comparisons). We incorporated Dale's law with 80% E, 20% I neurons into subsequent sweeps, **(b)** We varied the percentage of feedforward E-to-I connections. Minimal representations with chance color decode accuracy emerged when the percentage of feedforward E to I connections was 2% or less (feedforward E to E was fixed at 10%). **(c-d)** Color information was relatively robust to feedforward E-E connections and feedback connections. **(e)** At least 3 areas were required for the RNN's last area to resemble PMd dynamics. **(f)** 3-area RNNs with neurophysiological constraints had minimal representations that were generally robust to machine learning hyperparameters. The only exceptions were when the number of units was relatively small, or the learning rate was relatively large.

neural variance. However, it is possible there is significant color variance within a reach direction that dPCA, a linear dimensionality reduction technique, does not capture.

As we are interested in whether the color information is contained is the representation, a more appealing measure is decode accuracy. If the color of the target can be decoded from the representation of neural activity, then color information is present in the representation. We performed nonlinear dimensionality reduction via t-distributed stochastic neighbor embedding (tSNE) [39], shown in Fig. 2g. These results suggest that Areas 1 and 2 contain color information, but Area 3 does not (color decisions overlap). We decoded the color decision and context (target configuration) from RNN units in each area (Fig. 2h, Area 1, 2, and 3 color accuracy: 0.93, 0.76, 0.51, and Area 1, 2, and 3 context accuracy: 0.99, 0.87, 0.54). Area 1 and 2 had above chance context and color decode accuracies ($p < 0.01/9$, 1-tailed t-test with Bonferroni correction), while Area 3 color and context decode accuracies were near chance, and most similar to PMd (Fig. 2h, color: $p = 0.05$, context: $p = 0.024$). The direction decision could be decoded significantly above chance in all areas (Fig. 2h, $p < 0.01/9$). We also observed that Area 3 unit PSTHs more closely resembled PMd neuron PSTHs (e.g., Fig. 2i), and color CP progressively decreased in later areas (Fig. 2j). Area 3, like PMd, had many neurons with moderate to high direction CP, but low color CP.

We tested how robust these results were to architecture and hyperparameter selection [1] In particular, we quantified how well color could be decoded in the multi-area RNN's last area across several hyperparameter settings. We found that architecture impacted whether optimized multi-area RNNs had PMd-like minimal representations. In particular, we found that PMd-like representations emerged when we incorporated anatomical and neurophysiological constraints: Dale's law, empirical levels of feedforward inhibition, and at least 3 areas (Fig. 3a-e). When we varied machine learning hyperparameters, we found that our results were generally robust: multi-area RNNs had PMd-like representations in their last area over a wide range of hyperparameter settings (Fig. 3f). Together, this constellation of results shows that Area 3 of the multi-area RNN recapitulates key features of PMd activity, making this RNN a candidate model of multi-area decision-making in the Checkerboard Task.

[1] These sweeps over different random initializations and different parameter settings consisted of the most significant computational cost, roughly requiring 500 CPU hours on AWS, with each model training in approximately 1-2 hours.

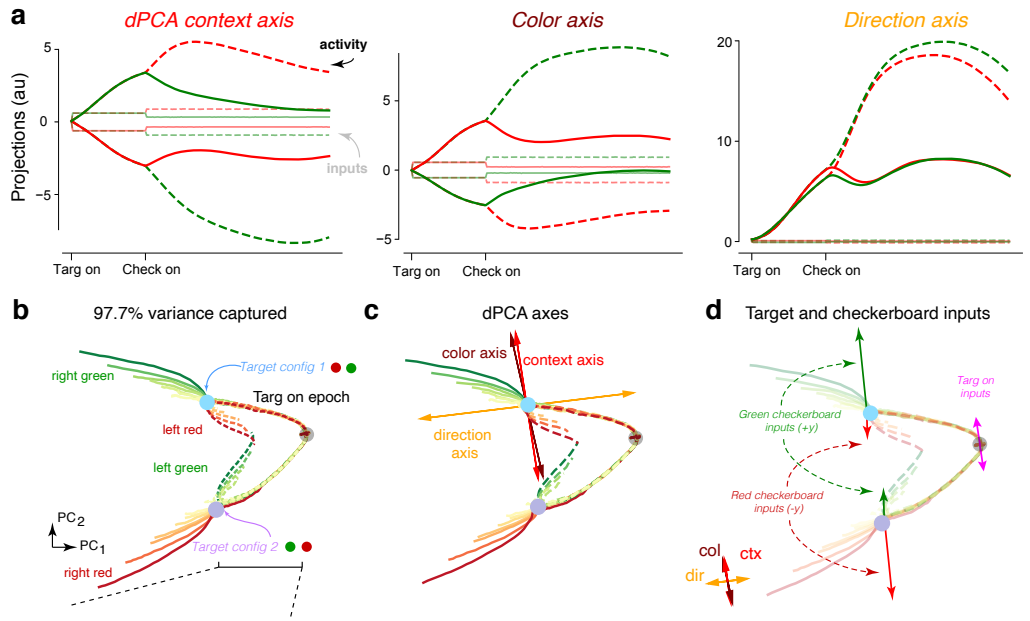

Figure 4: **Separation of direction and color in Area 1.** **(a)** The context, color, and direction axis correspond to the dPCA principal axes, which are not constrained to be orthogonal. Trajectories for different contexts and colors were separable on both the context and color axis. In contrast, the direction axis separated primarily on chosen direction. The RNN input representation had strong projections on the context and color axes, but not the direction axis. **(b)** Top 2 PCs of Area 1 activity, which capture 97.7% of the Area 1 variance. In the targets on epoch, the trajectories separate to two regions corresponding to the two potential target configurations (Target config 1 in blue, and Target config 2 in purple). The trajectories separate upon checkerboard color input, leading to four total trajectory motifs: right green, left red, right red, and left green. **(c)** Projection of the dPCA principal axes onto the PCs. **(d)** Projection of the context and color inputs onto the PCs. Context inputs are shown in pink, a green checkerboard input in green, and a red checkerboard input in red. Green (red) checkerboards lead to an increase (decrease) in PC_2 and the color axis, and differ in magnitude depending on the location of the trajectory in PC space. Trajectories are reduced in opacity to better visualize inputs.

In the next sections, we leverage the full observability of this biologically-plausible multi-area RNN to understand the mechanisms in different areas of the network and also how the network filters color information while propagating direction information.

3.2 Separation of the color and direction decision in Area 1

What are the key computational features of how the multi-area RNN represents color and direction information in the Checkerboard task? We first focused our analysis on Area 1, which uniquely has substantial variance for both color and direction decisions (Fig. 2h), implying a central role in computing the direction choice. We performed dPCA to identify demixed principal components related to the RNN inputs (coherence and context) and decisions (color and direction) [32]. We found demixed components that separated information related to coherence, context, the color choice, and the direction choice (Fig. 4b), consistent with these quantities being decodable from activity (Fig. 2h). We subsequently identified the context, color, and direction axes as the dPCA principal axes (unit norm, analogous to PCA eigenvectors), which combine the demixed components (analogous to PCA scores) to reconstruct neural activity [32].

We projected RNN activity and input representations onto the principal axes for context, color, and direction (Fig. 4a). We found that the context and color axis both responded to context and color inputs, and overall trajectories represented both context and color information. This suggests that color and context information are mixed in Area 1. In contrast, the direction axis strongly represented

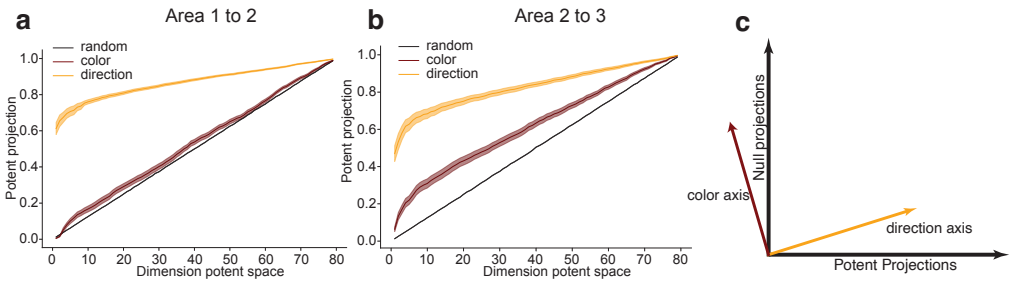

Figure 5: **(a)** Projections onto the potent space between Areas 1 and 2 for the color and direction axis, and a random vector as a function of effective rank for the input area to the middle area. Regardless of the dimension of the potent space, the direction axis is preferentially aligned with the potent space, indicating the information along this axis propagates, while the color axis is more aligned with the nullspace. Shading indicates s.e.m. **(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.

the direction choice, but did not strongly represent context or color (Fig. 4a, right). Strikingly, context and color inputs had nearly zero projection on the direction axis (Fig. 4a, right, opaque traces at 0). Consistent with these observations, we found the color and context axes were highly overlapping (dot product: 0.93), indicating that context and checkerboard variance are mixed in Area 1 activity. In contrast, the direction axes was closer to orthogonal to the context and color axes (overlap with color and context: 0.14 and 0.09, respectively).

These conclusions were upheld when we performed targeted dimensionality reduction (TDR), where we found (1) a direction axis separating left and right choices, with negligible input projections, and (2) that color and context representations were mixed (Appendix Fig. 14). Further, this structure was unique for PMd-like 3-area RNNs. In single-area RNNs, dPCA identified nearly orthogonal context, color, and direction axes, with trajectories that separated almost exclusively based on context, color, and direction, respectively (Appendix Fig. 13a).

This Area 1 representation has an important property: the direction choice is represented robustly on a nearly orthogonal axis that has close to zero context and color input projections, (Fig. 4e). This is not trivial: as counter-examples, single-area RNNs use direction axes that have context and color input projections (Fig. 13a), while the direction axis of an unconstrained 3-area RNN (without anatomical connectivity constraints that did not resemble PMd) has context and color information, and also receives context and color inputs (Fig. 13b). We show the axes overlapped with the PCs in Fig.4, as well as the effect of the checkerboard and target inputs, which qualitatively shows that the inputs do not project onto the direction axis.

3.3 Inter-area connections preferentially propagate output-relevant direction information

The differentiating aspect of multi-area computation is that the different areas are separated. A natural question to ask is how then does information propagate between areas? We denote the feedforward connections from Area 1 to 2 as $\mathbf{W}_{12}$, and from Area 2 to 3 as $\mathbf{W}_{23}$. We present results for feedforward connections from excitatory connections to excitatory units. Based on the hypothesis that the brain uses null and potent spaces to selectively filter information [40], we evaluated the effective potent and null spaces of $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$. We defined the effective potent space to be the right singular vectors corresponding to the largest singular values (see Methods). The effective null space corresponded to the singular vectors with the smallest singular values.

We quantified 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. 5a,b for $\mathbf{W}_{12}$ and $\mathbf{W}_{23}$, respectively. The null projection magnitudes are equal to one minus the potent projection. We found the direction axis was more aligned with the potent space and the color axis was more aligned with the null space. In fact, the direction axis (computed using the activity in Area 1) was consistently most aligned with the top singular vector (governed by the parameters of the feedforward matrix; which do not affect the activity in Area 1). In contrast, the color axis was similarly aligned to a random vector. This alignment was robust to the dimension of the effective potent space, and was

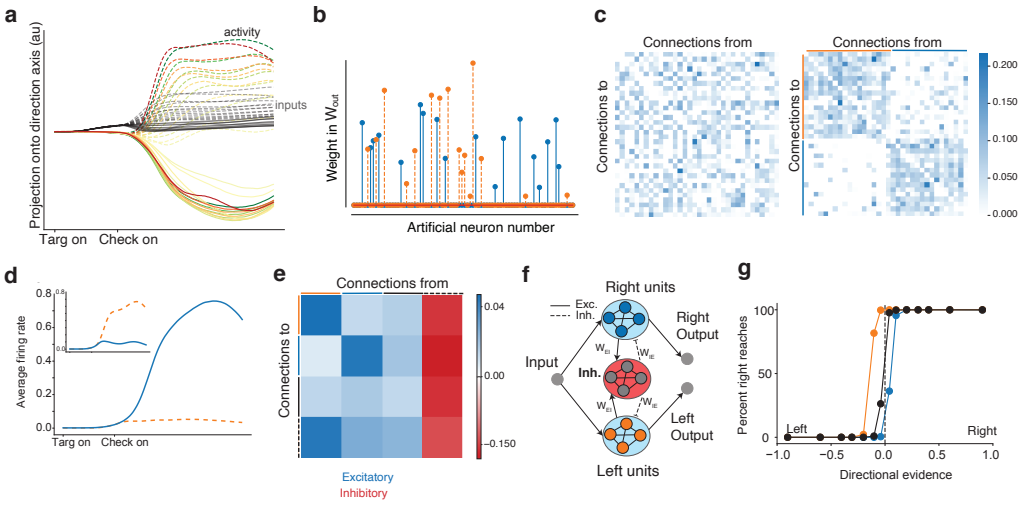

Figure 6: **Area 3 mechanism.** **(a)** Projection of input and overall activity onto the direction axis identified through dPCA. The conventions are the same as in Figure 4. **(b)** Readout weights in $\mathbf{W}_{\text{out}}$ are sparse, with many zero entries, and selective weights for a left or right 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. When one pool increases activity, the other pool decreases activity. **(e)** Averaged recurrent connectivity matrix. **(f)** Schematic of output area. **(g)** Psychometric curve after perturbation experiment, where 10% 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. Increasing inhibition to the left excitatory pool leads to more right choices and vice versa.

consistent across networks with varying feedforward connectivity percentages (10%, 20%, 30%, 50%, 100%). This suggests that learning in the multi-area recurrent network involved aligning the relevant information (in the activations) with the top singular vector (governed by the learned parameters of the feedforward matrix). These results indicate that direction information is preferentially propagated to subsequent areas, while color information is not. This phenomena is schematized in Fig. 5c.

These results also have implications on how inter-area connections relay information between areas. Color activity has significant representation in Area 1 (see Fig. 2). Therefore, the inter-area connections must not merely propagate the highest variance dimensions of a preceding area [41]. Consistent with this reasoning, we found that while the top 2 PCs capture 97.7% excitatory unit variance, the top 2 readout dimensions of $\mathbf{W}_{12}$ only captured 40.0% of Area 1's excitatory unit neural variance (Appendix Fig. 16). Hence, inter-area connections are not aligned with the most variable dimensions, but are rather aligned to preferentially propagate certain types of information — a result consistent with a recent study analyzing links between activity in V1 and V2 [41].

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

We showed previously that Area 3 most closely resembled PMd's dynamics (Fig. 2). Our results suggest that a direction signal has been computed before Area 3 and is selectively propagated through the RNN's inter-area connections. We found that the input to Area 3 (through $\mathbf{W}_{23}$) is a graded direction signal that provides a directional evidence signal for left or right reaches (Fig. 6a). This activity must be transformed into eventual DV outputs, which are the accumulated evidence for a left or right reach. This is illustrated in Fig. 6a, where we plot $\mathbf{W}_{23}\mathbf{r}_t^2$ ($\mathbf{r}_t^2$ are the unit activations of Area 2), and $\mathbf{r}_t^3$.

To analyze Area 3's dynamics, we first observed that $\mathbf{W}_{\text{out}}$'s coefficients were sparse, with 44 out of 80 output weights being identically zero. We found that the readout led to two separate clusters

of artificial units: units with non-zero coefficients for the left DV (orange) and those with non-zero coefficients for the right DV (blue). Artificial units projected either 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 and visualized the excitatory units of $\mathbf{W}_{\text{rec}}$, which upon first glance generally has no discernible structure (Fig. 6c, left panel). After sorting, we found that two self-excitatory pools of units emerged in $\mathbf{W}_{\text{rec}}$, the first pool in Fig. 6c (right) corresponding to the left DV and the second pool corresponding to the right DV. 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 Appendix Fig. 17. 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 (Fig. 6d). By taking the averaged connectivity matrix, similar to [42], we confirmed that there were two excitatory pools that received similar projections from the random excitatory pool and inhibitory pool (Fig. 6e). We summarize the behavior with a schematic of the area in Fig. 6f.

We subsequently applied selective perturbations to $\mathbf{W}_{\text{rec}}$ to determine how behavioral performance was biased. We increased inhibition to either the right or the left pool by doubling the weights of 10% of the inhibitory neurons associated with each pool. We found that this biased the network towards more left or right reaches, respectively, shown in Fig. 6g. 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 direction inputs to a decision variable through a winner-take-all mechanism.

4 Discussion

Even though behavior and cognition arise from the coordinated computations of multiple brain areas, there is limited understanding of how interacting brain areas coordinate to produce cognitive behavior [41, 43]. In this study, we used multi-area RNNs to gain mechanistic insight into how the brain computes a perceptual decision in the Checkerboard Task and transmits only the direction decision to PMd. These results propose hypotheses for computations that occur upstream of PMd, particularly how neural population activity representing context, color, and direction are structured, and what information is propagated between areas. We found that inter-area connections were preferentially aligned to the direction axis, not axes of maximal variance, leading to selective propagation of direction activity and attenuation of color activity. This role for inter-area connections is consistent with null and potent spaces for filtering and propagating information between areas [40, 44] and communication subspaces, which are aligned with lower variance dimensions [41].

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 or discarded. In the Checkerboard Task, color information is necessary to compute the direction decision, but does not need to be represented after the direction decision is computed, as in PMd [4, 26, 30, 31, 45]. In deep neural networks, it is believed that minimal representations simplify the role of the output classifier [46, 47]. This idea is consistent with (1) the multi-area RNN developing a minimal (little color information) but sufficient (robust direction information) representation of task inputs, and (2) Area 3, the output area, using a simple winner-take-all readout, forming two pools of neurons representing right and left decisions (Fig. 6).

References

[1] S. M. Szczepanski, C. S. Konen, and S. Kastner, "Mechanisms of spatial attention control in frontal and parietal cortex," *Journal of Neuroscience*, vol. 30, no. 1, pp. 148–160, 2010. [Online]. Available: https://www.jneurosci.org/content/30/1/148

[2] J. Kalaska and D. Crammond, "Cerebral cortical mechanisms of reaching movements," *Science*, vol. 255, no. 5051, pp. 1517–1523, 1992. [Online]. Available: https://science.sciencemag.org/content/255/5051/1517

[3] E. D. Remington, D. Narain, E. A. Hosseini, and M. Jazayeri, "Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics," *Neuron*, vol. 98, no. 5, pp. 1005–1019.e5, Jun. 2018.

[4] T. Yamagata, Y. Nakayama, J. Tanji, and E. Hoshi, "Distinct information representation and processing for goal-directed behavior in the dorsolateral and ventrolateral prefrontal cortex and the dorsal premotor cortex," *Journal of Neuroscience*, vol. 32, no. 37, pp. 12 934–12 949, 2012.

[5] J. M. Yau, G. C. DeAngelis, and D. E. Angelaki, "Dissecting neural circuits for multisensory integration and crossmodal processing," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1677, p. 20140203, 2015. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2014.0203

[6] P. Cisek, "Making decisions through a distributed consensus," *Current opinion in neurobiology*, vol. 22, no. 6, pp. 927–936, 2012.

[7] L. Pinto, K. Rajan, B. DePasquale, S. Y. Thiberge, D. W. Tank, and C. D. Brody, "Task-dependent changes in the large-scale dynamics and necessity of cortical regions," *Neuron*, vol. 104, no. 4, pp. 810–824, 2019.

[8] M. Siegel, T. J. Buschman, and E. K. Miller, "Cortical information flow during flexible sensorimotor decisions," *Science*, vol. 348, no. 6241, pp. 1352–1355, 2015.

[9] D. J. Freedman and J. A. Assad, "Neuronal mechanisms of visual categorization: an abstract view on decision making," *Annual review of neuroscience*, vol. 39, pp. 129–147, 2016.

[10] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 23, pp. 8619–8624, Jun. 2014.

[11] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nat. Neurosci.*, vol. 19, no. 3, pp. 356–365, Mar. 2016.

[12] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *Nature*, vol. 503, no. 7474, pp. 78–84, Nov. 2013.

[13] R. Laje and D. V. Buonomano, "Robust timing and motor patterns by taming chaos in recurrent neural networks," *Nature neuroscience*, vol. 16, no. 7, p. 925, 2013.

[14] A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil *et al.*, "Vector-based navigation using grid-like representations in artificial agents," *Nature*, vol. 557, no. 7705, pp. 429–433, 2018.

[15] G. Hennequin, T. P. Vogels, and W. Gerstner, "Optimal control of transient dynamics in balanced networks supports generation of complex movements," *Neuron*, vol. 82, no. 6, pp. 1394–1406, Jun. 2014.

[16] J. P. Stroud, M. A. Porter, G. Hennequin, and T. P. Vogels, "Motor primitives in space and time via targeted gain modulation in cortical networks," *Nature neuroscience*, vol. 21, no. 12, p. 1774, 2018.

[17] O. Barak, "Recurrent neural networks as versatile tools of neuroscience research," *Current Opinion in Neurobiology*, vol. 46, pp. 1 – 6, 2017, computational Neuroscience. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438817300429

[18] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy, "A neural network that finds a naturalistic solution for the production of muscle activity," *Nat. Neurosci.*, vol. 18, no. 7, pp. 1025–1033, 2015.

[19] H. F. Song, G. R. Yang, and X. J. Wang, "Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework," *PLoS Comput. Biol.*, vol. 12, no. 2, pp. 1–30, 2016.

[20] H. F. Song, G. R. Yang, and X.-J. Wang, "Reward-based training of recurrent neural networks for cognitive and value-based tasks," *Elife*, vol. 6, Jan. 2017.

[21] W. Chaisangmongkon, S. K. Swaminathan, D. J. Freedman, and X. J. Wang, "Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions," *Neuron*, vol. 93, no. 6, pp. 1504–1517.e4, 2017.

[22] F. Mastrogiuseppe and S. Ostojic, "Linking connectivity, dynamics, and computations in low-rank recurrent neural networks," *Neuron*, vol. 99, no. 3, pp. 609 – 623.e29, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627318305439

[23] D. Sussillo and O. Barak, "Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks," *Neural Comput.*, vol. 25, no. 3, pp. 626–649, 2013.

[24] C. Chandrasekaran, D. Peixoto, W. T. Newsome, and K. V. Shenoy, "Laminar differences in decision-related neural activity in dorsal premotor cortex," *Nat. Commun.*, vol. 8, no. 1, p. 614, Sep. 2017.

[25] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex," *Elife*, vol. 4, pp. 1–21, 2015.

[26] É. Coallier, T. Michelet, and J. F. Kalaska, "Dorsal premotor cortex: neural correlates of reach target decisions based on a color-location matching rule and conflicting sensory evidence," *Journal of neurophysiology*, vol. 113, no. 10, pp. 3543–3573, 2015.

[27] J. I. Gold and M. N. Shadlen, "The neural basis of decision making," *Annual review of neuroscience*, vol. 30, 2007.

[28] R. Romo and V. de Lafuente, "Conversion of sensory signals into perceptual decisions," *Progress in Neurobiology*, vol. 103, pp. 41 – 75, 2013, conversion of Sensory Signals into Perceptions, Memories and Decisions. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301008212000391

[29] D. J. Freedman and J. A. Assad, "A proposed common neural mechanism for categorization and perceptual decisions," *Nature neuroscience*, vol. 14, no. 2, p. 143, 2011.

[30] C. Chandrasekaran, D. Peixoto, W. T. Newsome, and K. V. Shenoy, "Laminar differences in decision-related neural activity in dorsal premotor cortex," *Nature Communications*, vol. 8, no. 1, p. 614, 2017.

[31] M. Wang, C. Montanède, C. Chandrasekaran, D. Peixoto, K. V. Shenoy, and J. F. Kalaska, "Macaque dorsal premotor cortex exhibits decision-related activity only when specific stimulus-response associations are known," *Nature Communications*, vol. 10, no. 1, p. 1793, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-09460-y

[32] D. Kobak, W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X.-L. Qi, R. Romo, N. Uchida, and C. K. Machens, "Demixed principal component analysis of neural population data," *Elife*, vol. 5, Apr. 2016.

[33] S. P. Wise and E. A. Murray, "Arbitrary associations between antecedents and actions," *Trends in Neurosciences*, vol. 23, no. 6, pp. 271 – 276, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166223600015708

[34] E. Hoshi, "Cortico-basal ganglia networks subserving goal-directed behavior mediated by conditional visuo-goal association," *Front. Neural Circuits*, vol. 7, no. October, p. 158, 2013.

[35] D. Badre, J. Hoffman, J. W. Cooney, and M. D'esposito, "Hierarchical cognitive control deficits following damage to the human frontal lobe," *Nature neuroscience*, vol. 12, no. 4, p. 515, 2009.

[36] H. F. Song, G. R. Yang, and X. J. Wang, "Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework," *PLoS Comput. Biol.*, vol. 12, no. 2, pp. 1–30, 2016.

[37] N. T. Markov, M. M. Ercsey-Ravasz, A. R. Ribeiro Gomes, C. Lamy, L. Magrou, J. Vezoli, P. Misery, A. Falchier, R. Quilodran, M. A. Gariel, J. Sallet, R. Gamanut, C. Huissoud, S. Clavagnier, P. Giroud, D. Sappey-Marinier, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy, "A weighted and directed interareal connectivity matrix for macaque cerebral cortex," *Cereb. Cortex*, vol. 24, no. 1, pp. 17–36, Jan. 2014.

[38] K. W. Latimer, "Nonlinear demixed component analysis for neural population data as a low-rank kernel regression problem," *Neurons, Behavior, Data Analysis, & Theory*, pp. 1 – 24, 2019.

[39] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[40] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Cortical activity in the null space: permitting preparation without movement," *Nat. Neurosci.*, vol. 17, no. 3, pp. 440–448, Mar. 2014.

[41] J. D. Semedo, A. Zandvakili, C. K. Machens, B. M. Yu, and A. Kohn, "Cortical Areas Interact through a Communication Subspace," *Neuron*, vol. 102, no. 1, pp. 249–259.e4, apr 2019. [Online]. Available: https://doi.org/10.1016/j.neuron.2019.01.026

[42] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, "Task representations in neural networks trained to perform many cognitive tasks," *Nature neuroscience*, vol. 22, no. 2, pp. 297–306, 2019.

[43] A. Kohn, A. I. Jasper, J. D. Semedo, E. Gokcen, C. K. Machens, and B. M. Yu, "Principles of corticocortical communication: Proposed schemes and design considerations," *Trends in Neurosciences*, vol. 43, no. 9, pp. 725–737, 2020/10/12 2020.

[44] S. D. Stavisky, J. C. Kao, S. I. Ryu, and K. V. Shenoy, "Motor cortical visuomotor feedback activity is initially isolated from downstream targets in Output-Null neural state space dimensions," *Neuron*, vol. 95, no. 1, pp. 195–208.e9, Jul. 2017.

[45] P. Cisek and J. F. Kalaska, "Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action," *Neuron*, vol. 45, no. 5, pp. 801–814, 2005.

[46] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *Journal of Machine Learning Research*, vol. 19, no. 50, pp. 1–34, 2018. [Online]. Available: http://jmlr.org/papers/v19/17-646.html

[47] S. Soatto and A. Chiuso, "Visual representations: Defining properties and deep approximation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, June 2016.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[49] H. Barbas, "General cortical and special prefrontal connections: principles from structure to function," *Annual review of neuroscience*, vol. 38, pp. 269–289, 2015.

[50] D. Datta and A. F. T. Arnsten, "Loss of prefrontal cortical higher cognition with uncontrollable stress: Molecular mechanisms, changes with age, and relevance to treatment," *Brain Sci*, vol. 9, no. 5, May 2019.

[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, Nov. 2016.

[52] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy, "A neural network that finds a naturalistic solution for the production of muscle activity," *Nat. Neurosci.*, vol. 18, no. 7, pp. 1025–1033, 2015.

[53] J. A. Michaels, B. Dann, and H. Scherberger, "Neural population dynamics during reaching are better explained by a dynamical system than representational tuning," *PLoS Comput. Biol.*, vol. 12, no. 11, p. e1005175, 2016.

[54] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, Feb. 2013, pp. 1310–1318.

[55] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *Nature*, vol. 503, no. 7474, pp. 78–84, Nov. 2013.

[56] B. W. Brunton, M. M. Botvinick, and C. D. Brody, "Rats and humans can optimally accumulate evidence for decision-making," *Science*, vol. 340, no. 6128, pp. 95–98, 2013.

[57] R. Ratcliff, "A theory of memory retrieval." *Psychological review*, vol. 85, no. 2, p. 59, 1978.

[58] J. C. Kao, "Considerations in using recurrent neural networks to probe neural dynamics," *J. Neurophysiol.*, vol. 122, no. 6, pp. 2504–2521, Dec. 2019.

[59] D. Barber and F. Agakov, "The im algorithm: A variational approach to information maximization," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, ser. NIPS'03. Cambridge, MA, USA: MIT Press, 2003, p. 201–208.

[60] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 5171–5180. [Online]. Available: http://proceedings.mlr.press/v97/poole19a.html

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mention that the mechanisms in the RNNs should be interpreted as hypotheses for biological mechanisms
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not foresee any potential negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All details required to implement our experiments are described in the Appendix, and we have provided the code to our implementation.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See footnote on page 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]