

1 Clustering and compositionality of task representations in a neural  
2 network trained to perform many cognitive tasks

3 Guangyu Robert Yang<sup>1</sup>, H. Francis Song<sup>1,5</sup>, William T. Newsome<sup>3,4</sup>, Xiao-Jing Wang<sup>1,2</sup>

4 August 27, 2017

5 **Author Affiliations**

6 <sup>1</sup> Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003

7 <sup>2</sup> NYU-ECNU Institute of Brain and Cognitive Science, NYU Shanghai, Shanghai, China

8 <sup>3</sup> Department of Neurobiology, Stanford University, Stanford, CA 94305, USA

9 <sup>4</sup> Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA

10 <sup>5</sup> Current address: DeepMind, London EC4A 3TW, United Kingdom.

11 **Contact**

12 Xiao-Jing Wang: [xjwang@nyu.edu](mailto:xjwang@nyu.edu)

## 13 **ABSTRACT**

14 A neural system has the ability to flexibly perform many tasks, but the underlying mechanism  
15 cannot be elucidated in traditional experimental and modeling studies designed for one task at  
16 a time. Here, we trained a single network model to perform 20 cognitive tasks that may involve  
17 working memory, decision-making, categorization and inhibitory control. We found that after  
18 training, recurrent units developed into clusters that are functionally specialized for various cog-  
19 nitive processes. We introduce a measure to quantify relationships between single-unit neural  
20 representations of tasks, and report five distinct types of such relationships that can be tested  
21 experimentally. Surprisingly, our network developed compositionality of task representations, a  
22 critical feature for cognitive flexibility, whereby one task can be performed by recombining in-  
23 structions for other tasks. Finally, we demonstrate how the network could learn multiple tasks  
24 sequentially. This work provides a computational platform to investigate neural representations  
25 of many cognitive tasks.

## 26 **INTRODUCTION**

27 The prefrontal cortex is important for numerous cognitive functions<sup>1-3</sup>, partly because of its  
28 central role in task representation<sup>4-7</sup>. Electrophysiological experiments using behaving animals  
29 reported prefrontal neurons that are either selective for different aspects of a given task<sup>8,9</sup> or  
30 functionally mixed<sup>10,11</sup>. Much less is known about functional specialization of task representa-  
31 tions at the neuronal level. Imagine a single-neuron recording that could be carried out with  
32 animals switching between many different tasks. Is each task supported by a "private" set of  
33 neurons, or does each task involve every neuron in the network, or somewhere in between? If  
34 two tasks require a common underlying cognitive process, such as working memory or decision  
35 making, what would be the relationship between their neural representations? In other words,  
36 what would be the "neural relationship" between this pair of tasks? Would the two tasks utilize a  
37 shared neural substrate?

38 Humans readily learn to perform many cognitive tasks in a short time. By following verbal in-  
39 structions such as "Release the lever only if the second item is not the same as the first," humans  
40 can perform a novel task without any training at all<sup>6</sup>. A cognitive task is typically composed of el-

41 elementary sensory, cognitive, and motor processes<sup>5</sup>. Performing a task without training requires  
42 composing elementary processes that are already learned into temporal sequences that enable  
43 correct performance on the new task. This property, called "compositionality," has been pro-  
44 posed as a fundamental principle underlying flexible cognitive control<sup>12</sup>. Indeed, human studies  
45 have suggested that the representation of complex cognitive tasks in the lateral prefrontal cortex  
46 is compositional<sup>6,13</sup>. However, these tasks involved verbal instructions; it is unknown whether  
47 non-verbal tasks commonly used in animal physiological experiments also display composition-  
48 ality and whether relatively simple neural network models are sufficient to support composi-  
49 tional task structures.

50 These questions remain difficult to address with conventional experimental and modeling  
51 approaches. Experiments with laboratory animals have so far been largely limited to a single  
52 task at a time; on the other hand, human imaging studies lack the spatial resolution to address  
53 questions at the single neuron level. Therefore, the lack of neural recordings from animals per-  
54 forming many different tasks leaves unanswered important questions regarding how a single  
55 network represents and supports distinct tasks. In principle, these questions could be addressed  
56 in neural circuit models, but designing a single neural circuit model capable of multiple tasks is  
57 challenging and virtually nonexistent. To tackle these problems, we took the approach of train-  
58 ing recurrent neural networks (RNNs)<sup>11,14-19</sup>. In this work, we trained a single RNN to perform  
59 20 cognitive tasks. We found that after training, the emerging task representations are organized  
60 in the form of clustering of recurrent units. Our network also makes numerous testable predic-  
61 tions regarding the relationship between the neural representations of pairs of cognitive tasks.  
62 Surprisingly, we found that compositionality of task representations emerges from training in  
63 our network model, which can be instructed to perform new tasks without further training. Our  
64 work provides a framework for investigating neural representations of task structures and neural  
65 relationships between tasks.

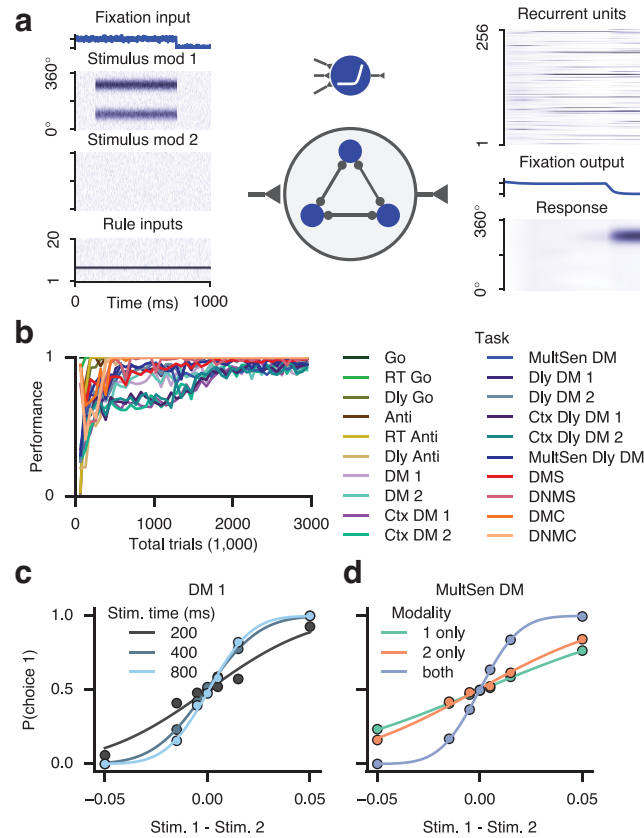
## 66 RESULTS

### 67 Training neural networks for many cognitive tasks

68 To study how various cognitive tasks might be implemented in a single neural circuit, we trained  
69 a recurrent neural network model (**Fig. 1a**) to perform 20 tasks, most of which are commonly  
70 used in neurophysiological studies of nonhuman animals and crucial to our understanding of  
71 the neural mechanisms of cognition. The chosen set of tasks includes variants of memory-  
72 guided response<sup>20</sup>, simple perceptual decision making<sup>21</sup>, context-dependent decision-making<sup>11,22</sup>,  
73 multi-sensory integration<sup>23</sup>, parametric working memory<sup>24</sup>, inhibitory control (e.g., in anti-saccade)<sup>25</sup>,  
74 delayed match-to-sample<sup>26</sup>, and delayed match-to-category<sup>27</sup> tasks (**Table 1, Supplementary**  
75 **Fig. 1**).

76 The recurrent network model emulates a "cognitive-type" cortical circuit such as the pre-  
77 frontal cortex<sup>3</sup>, which receives converging inputs from multiple sensory pathways and projects  
78 to downstream motor areas. We designed our network architecture to be general enough for all  
79 the tasks mentioned above, but otherwise as simple as possible to facilitate analysis. For every  
80 task, the network receives noisy inputs of three types: fixation, stimulus, and rule (**Fig. 1a**). The  
81 fixation input indicates whether the network should "fixate" or respond (e.g. "saccade"). Thus  
82 the decrease in the fixation input provides a "go signal" to the network. The stimulus inputs con-  
83 sist of two modalities, each represented by a ring of input units that encodes a one-dimensional  
84 circular variable such as motion direction or color on a color wheel<sup>18</sup>. A single rule input unit is  
85 activated in each trial, instructing the network on which task it is currently supposed to perform.  
86 The network projects to a fixation output unit and a group of motor units encoding the response  
87 direction as a one dimensional variable on a ring of outputs (e.g., saccade direction, reach direc-  
88 tion). To mimic biological neurons, all units in our recurrent network receive private noise and  
89 have non-negative activities, imposed by a realistic neuronal input-output function<sup>28</sup>.

90 Before training, a network is incapable of performing any task. It is trained with supervised  
91 learning<sup>11,15</sup>, which modifies all connection weights (input, recurrent, and output) to minimize  
92 the difference between the network output and a desired (target) output. All tasks were randomly  
93 interleaved during training (at the end we will present results from sequential training). Below  
94 we show results obtained from networks of 256 recurrent units, and results are robust with re-



**Figure 1.** A recurrent neural network model is trained to perform a large number of cognitive tasks. **(a)** A recurrent neural network (middle) described by rate units receives inputs (left) encoding a fixation cue, stimuli from two modalities, and a rule signal (which instructs the system which task to perform in a given trial). The network has 256 recurrent units (top right), and it projects to a fixation output unit (which should be active when a motor response is unwarranted) and a population of units selective for response directions (right). All units in the recurrent network have non-negative firing rates. All connection weights and biases are modifiable by training using a supervised learning protocol. **(b)** The network successfully learned to perform 20 tasks. **(c, d)** Psychometric curves in two decision making (DM) tasks. **(c)** Perceptual decision-making relies on temporal integration of information, as the network performance improves when the noisy stimulus is presented for a longer time. **(d)** In a multi-sensory integration task, the trained network combines information from two modalities to improve performance (compared with performance when information is only provided by a single modality).

| Task name                                   | Abbreviation   | Task family | Reference |
|---|----------------|-------------|-----------|
| Go  | Go             | Go          | N/A       |
| Reaction-time go                            | RT Go          | Go          | N/A       |
| Delayed go                                  | Dly Go         | Go          | [20]      |
| Anti-response                               | Anti           | Anti        | [25]      |
| Reaction-time anti-response                 | RT Anti        | Anti        | [25]      |
| Delayed anti-response                       | Dly Anti       | Anti        | [25]      |
| Decision making 1                           | DM 1           | DM          | [21]      |
| Decision making 2                           | DM 2           | DM          | [21]      |
| Context-dependent decision making 1         | Ctx DM 1       | DM          | [11]      |
| Context-dependent decision making 2         | Ctx DM 2       | DM          | [11]      |
| Multi-sensory decision making               | MultSen DM     | DM          | [23]      |
| Delayed decision making 1                   | Dly DM 1       | Dly DM      | [24]      |
| Delayed decision making 2                   | Dly DM 2       | Dly DM      | [24]      |
| Context-dependent delayed decision making 1 | Ctx Dly DM 1   | Dly DM      | N/A       |
| Context-dependent delayed decision making 2 | Ctx Dly DM 2   | Dly DM      | N/A       |
| Multi-sensory delayed decision making       | MultSen Dly DM | Dly DM      | N/A       |
| Delayed match-to-sample                     | DMS            | Matching    | [26]      |
| Delayed non-match-to-sample                 | DNMS           | Matching    | [26]      |
| Delayed match-to-category                   | DMC            | Matching    | [27]      |
| Delayed non-match-to-category               | DNMC           | Matching    | [27]      |

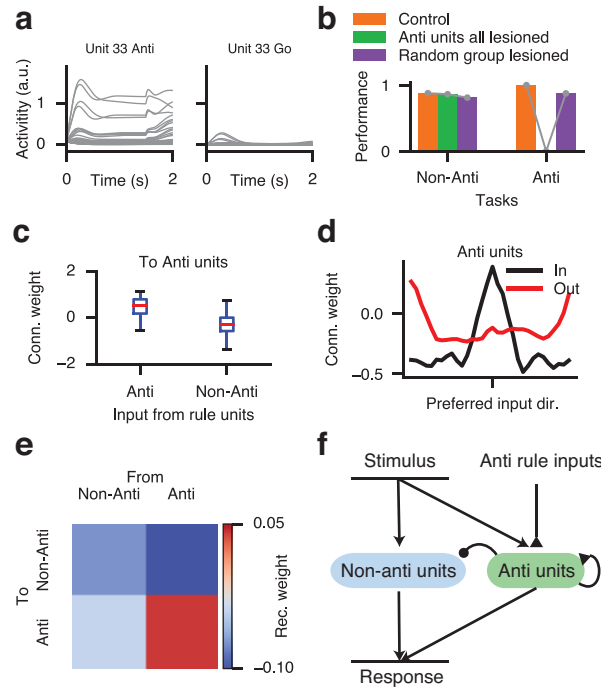
**Table 1.** Names and abbreviations of all tasks trained in the networks. Most of the trained tasks are derived from archetypal cognitive tasks used in non-human animal experiments. We grouped our tasks into five task families. We are not aware of experimental studies that investigated the Ctx Dly DM 1, Ctx Dly DM 2, or MultSen Dly DM tasks in non-human animals.

95 spect to the exact network size. After training, a single network model achieved high behavioral  
96 performance across all tasks (**Fig. 1b**). Furthermore, by conducting a battery of psychomet-  
97 ric tests, we demonstrate that the network displays behavioral features consistent with animal  
98 studies. For instance, in perceptual decision-making tasks, the network achieves better perfor-  
99 mance with higher coherence and longer duration of the stimulus (**Fig. 1c, Supplementary Fig.**  
100 **2a-f**)<sup>21</sup>, and it combines information from different sources to form decisions (**Fig. 1d**)<sup>23</sup>. In  
101 working memory tasks, the network can maintain information throughout a delay period of up  
102 to five seconds (**Supplementary Fig. 2g**)<sup>1,20,24</sup>.

### 103 **Dissecting the circuit for the family of Anti tasks**

104 For trained neural networks to be useful model systems for neuroscience, it is critical that we  
105 attempt to understand the circuit mechanism underlying the network computation<sup>29</sup>. Here we  
106 demonstrate how a trained network could be dissected and analyzed in a sample family of cog-  
107 nitive tasks. Anti-response tasks are important tools to investigate voluntary action and inhibitory  
108 control<sup>25</sup>. These tasks require an anti-response, in the opposite direction from the more com-  
109 mon pro-response towards a stimulus' location. Our set of tasks includes three tasks from the  
110 Anti task family (**Table 1**): the anti-response (Anti), reaction-time anti-response (RT Anti), and  
111 delayed anti-response (Dly Anti) tasks. We found that a subgroup of units emerged in a trained  
112 network, which we call Anti units (**Fig. 2a**). These units are primarily selective to stimuli in the  
113 Anti family of tasks. Inactivating or "lesioning" all Anti units at once resulted in a complete fail-  
114 ure in performing the family of tasks that require an anti-response, but had essentially no impact  
115 on the performance of the other tasks (**Fig. 2b**).

116 Since we have access to all the information of the trained network, we next investigated the  
117 connection weights of Anti units to understand their roles. Anti units receive positive connec-  
118 tion weights from the three rule input units representing Anti tasks (**Fig. 2c**), which explained  
119 why Anti units are only active during Anti tasks. Next, we studied the connection weights of Anti  
120 units with the stimulus-encoding input ring and the response-encoding output ring. For each  
121 Anti unit, the preferred input and output directions defined by the input and output connection  
122 weights are 180 degrees apart (**Fig. 2d**). These opposite preferred directions serve as the neu-  
123 ral substrate for vector inversion (anti-mapping) required by Anti tasks. Finally, the Anti units



**Figure 2.** Dissecting the circuit for a family of tasks. **(a)** An example Anti unit, which is primarily selective in the Anti-family of tasks. Different traces show neural activities across stimulus conditions within a task. **(b)** After lesioning all Anti units together (green), the network can no longer perform any of the Anti tasks, while performance for other tasks remain intact. Instead, lesioning the same number of randomly selected units had a minor impact on the performance. **(c)** Anti units receive strong positive connections from rule units representing the Anti tasks but negative connections from non-Anti rule units. The box shows the lower to upper quartiles of connection weights. The whiskers show the full range. **(d)** Average connections from input units to Anti units (black) and those onto output units (red) display opposite preferred directions, thereby vector conversion (from pro- to anti-response) is realized. Both input and output connections are sorted by each unit's preferred input direction, defined as the stimulus direction represented by the strongest-projecting input unit. **(e)** Network wiring architecture that emerged from training, in which Anti units excite themselves and strongly inhibit other units. **(f)** Circuit diagram summarizing the neural mechanism of the Anti-family tasks.



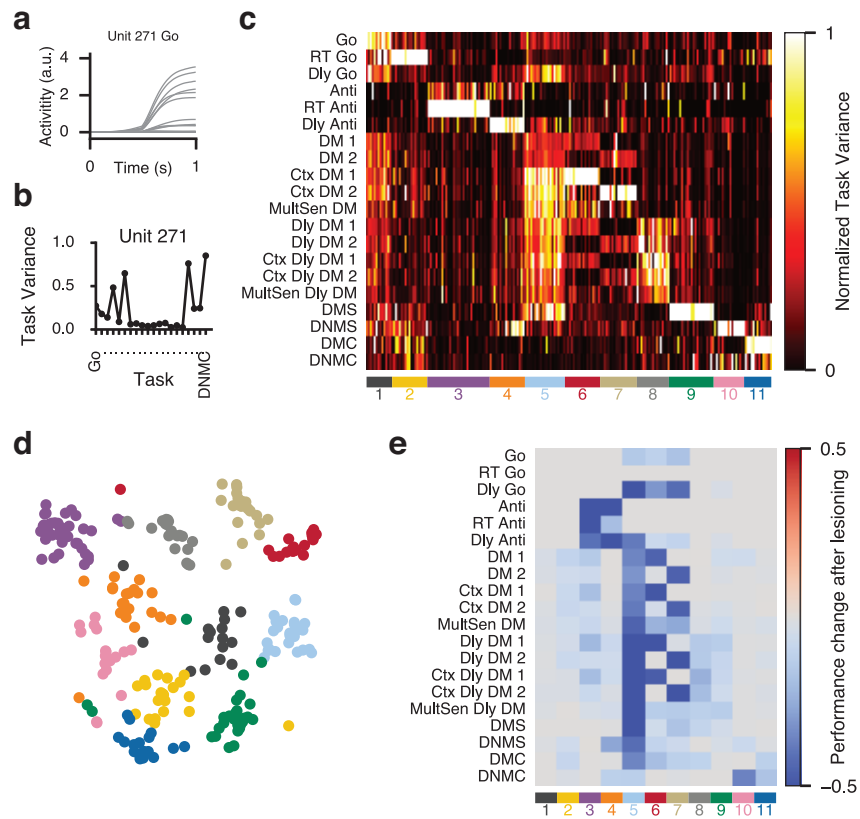
124 strongly inhibit the rest of the recurrent units (Non-Anti units) through recurrent connections  
125 (**Fig. 2e**), suppressing a pro-response with inhibitory control. Thus, the circuit mechanism un-  
126 derlying Anti tasks in our trained network is delineated: A group of units emerge from training  
127 that are specialized for the anti-response process and are essential in every task that requires this  
128 process. The Anti rule inputs engage vector-inverting Anti units, which in turn exert inhibitory  
129 control over Non-Anti units (**Fig. 2f**).

### 130 **Functional clusters encode subsets of tasks**

131 The focus of our analysis was to examine the neural representation of tasks. After training, it  
132 is conceivable that each unit of the recurrent network is only selective in one or a few tasks,  
133 forming highly-specialized task representations. On the other hand, task representations may  
134 be completely mixed, where all units are engaged in every task. We sought to assess where our  
135 network lies on the continuum between these two extreme scenarios.

136 To quantify single-unit task representation, we need a measure of task selectivity that is gen-  
137 eral enough so it applies to a broad range of tasks, and at the same time simple enough so it  
138 can be easily computed. We propose a measure that we call Task Variance (see Online Meth-  
139 ods). For each unit, the task variance for a given task is obtained by first computing the variance  
140 of neural activities across all possible task conditions at a given time point, then averaging that  
141 variance across time (excluding the fixation epoch) (**Fig. 3a**). Task variance is agnostic about  
142 the task setup and can be easily computed in models and is also applicable to the analysis of  
143 experimental data.

144 By computing the task variance for all trained tasks, we can study how individual units are  
145 differentially selective in all the tasks (**Fig. 3b**). For better comparison across units, we normal-  
146 ized the task variance of each unit such that the maximum normalized variance over all tasks is  
147 one. By analyzing the patterns of normalized task variance for all active units, we found that units  
148 are self-organized into distinct clusters through learning (**Fig. 3c,d**) (see Online Methods). We  
149 identified about 10 clusters in the network. Each cluster is mainly selective in a specific subset of  
150 tasks. To understand the causal role of these clusters, we lesioned each of them while monitoring  
151 the change in performance across all 20 tasks (**Fig. 3e**). We found one cluster (cluster number  
152 3) that is specialized for the Anti-family tasks, and it consists mainly of Anti units analyzed in



**Figure 3.** The emergence of functionally specialized clusters for task representation. (a) Neural activity of a single unit during an example task. Different traces correspond to different stimulus conditions. (b) Task variances across all tasks for the same unit. For each unit, task variance measures the variance of activities across all stimulus conditions. (c) Task variances across all tasks and active units, normalized by the peak value across tasks for each unit. Units form distinct clusters identified using the K-means clustering method based on normalized task variances. Each cluster is specialized for a subset of tasks. A task can involve units from several clusters. Units are sorted by their cluster membership, indicated by colored lines at the bottom. (d) Visualization of the task variance map. For each unit, task variances across tasks form a vector that is embedded in the two-dimensional space using t-distributed Stochastic Neighbor Embedding (t-SNE). Units are colored according to their cluster membership. (e) Change in performance across all tasks when each cluster of units is lesioned.

153 **Fig. 2.** Another two clusters (cluster numbers 6 and 7) are specialized for decision-making tasks  
154 involving modality 1 and 2 respectively. Furthermore, one cluster (cluster number 5) selective  
155 in the parametric working memory tasks (Dly DM task family) is also selective in the perceptual  
156 decision making tasks (DM task family), indicating a common neural substrate for these two  
157 cognitive functions in our network<sup>30</sup>. We can also study how units are clustered based on epoch  
158 variance, a measure that quantifies how selective units are in each task epoch (**Supplementary**  
159 **Fig. 3**). One cluster of units presumably supports response generation, as it is highly selective in  
160 the response epoch but not the stimulus epoch. Our results indicate that the network success-  
161 fully identified common sensory, cognitive, and motor processes underlying subsets of tasks,  
162 and through training developed units dedicated to the shared processes rather than the individ-  
163 ual tasks.

#### 164 **Relationships between neural representations of pairs of tasks**

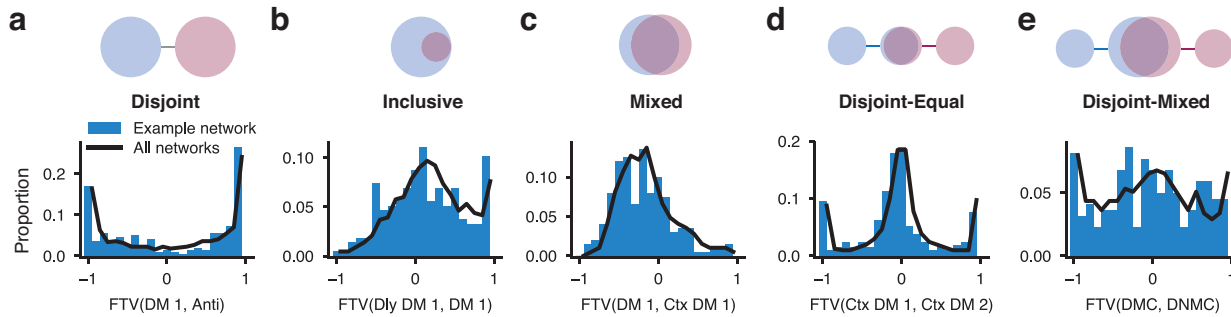
165 The map of normalized task variance in **Fig. 3c** allowed us to visualize the whole network across  
166 many tasks all at once. However, it is of limited use when we try to compare with experimental  
167 data or to analyze the (dis)similarity of the neural task representation between any pair of tasks.  
168 To quantify how each unit is selective in one task in comparison to another task, we introduce  
169 a simple measure based on task variance: the Fractional Task Variance (FTV). For unit  $i$ , the  
170 fractional task variance with respect to task A and task B is defined as

$$FTV_i(A, B) = \frac{TV_i(A) - TV_i(B)}{TV_i(A) + TV_i(B)}, \quad (1)$$

171 where  $TV_i(A)$  and  $TV_i(B)$  are the task variances for tasks A and B respectively. Fractional task  
172 variance ranges between  $-1$  and  $+1$ . Having a  $FTV_i(A, B)$  close to  $+1$  (or  $-1$ ) means that unit  $i$  is  
173 primarily selective in task A (or B).

174 For every pair of tasks, we can compute the fractional task variance for all units that are active  
175 in at least one of the two tasks. Each distribution of FTVs contains rich information about the  
176 single-unit level neural relationship between the pair of tasks. Having 20 tasks provides us with  
177 190 distinct FTV distributions (**Supplementary Fig. 4**), from the shape of which we summarized  
178 five typical neural relationships (**Fig. 4**).

179 1. Disjoint (**Fig. 4a**). When two tasks have a disjoint relationship like the Anti task and the



**Figure 4.** A diversity of neural relationships between pairs of tasks. For a pair of tasks, we characterize their neural relationship by the distribution of fractional task variances over all units. We observed five typical relationships: Disjoint (**a**), Inclusive (**b**), Mixed (**c**), Disjoint-Equal (**d**), and Disjoint-Mixed (**e**). Blue: distribution for one example network. Black: averaged distribution over 20 networks.

180 DM1 task, the FTV distribution is characterized by two peaks at the two ends and few units in  
181 between. There is little overlap between units selective in the two tasks. The shape of the FTV  
182 distribution is rather robust across independently trained networks: The FTV distribution from  
183 one sample network closely matches the averaged distribution from 20 networks.

184 2. Inclusive (**Fig. 4b**). This relationship is embodied by a strongly skewed FTV distribution,  
185 suggesting that one task is neurally a subset of another task. In this case, there are no units that  
186 are selective in the DM1 task yet not in the Dly DM 1 task.

187 3. Mixed (**Fig. 4c**). A mixed relationship is characterized by a broad uni-modal FTV distri-  
188 bution centered around 0 with no clear peak at the two ends. This distribution suggests that the  
189 two tasks utilize overlapping neural circuits.

190 4. Disjoint-Equal (**Fig. 4d**). For Ctx DM 1 and 2, the FTV distribution is trimodal, with two  
191 peaks at the two ends and an additional peak around 0. This relationship can be considered as  
192 a combination of the Disjoint relationship and the Equal relationship. The Equal relationship is  
193 represented by a single, narrow peak around 0. In this scenario, the two tasks each gets a private  
194 neural population, while they also share the third population.

195 5. Disjoint-Mixed (**Fig. 4e**). This relationship is a combination of the Disjoint and the Mixed  
196 relationships. Many units only participate in one of the two tasks, while the rest of the units are  
197 mixed in both tasks.

198 In summary, we introduced a simple yet informative measure to study the diverse neural

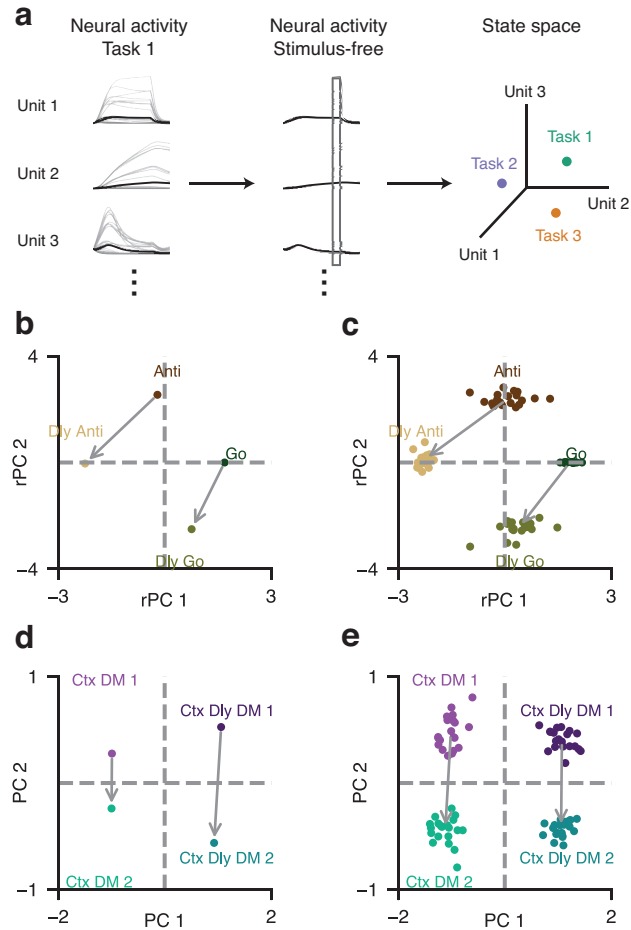
199 relationships between pairs of tasks. We found that these relationships can be categorized into  
200 several canonical types. Our results on FTV distributions (**Supplementary Fig. 4**) provide an  
201 array of straightforward predictions on pairwise neural relationships between cognitive tasks.

## 202 **Compositional representations of tasks**

203 A cognitive task can, in general, be expressed abstractly as a sequence of sensory, cognitive and  
204 motor processes, and cognitive processes may involve a combination of basic functions (such  
205 as working memory) required to perform the task. The compositionality of cognitive tasks is  
206 natural for human subjects because tasks are instructed with natural languages, which are com-  
207 positional in nature<sup>12</sup>. For example, the Go task can be instructed as "Saccade to the direction  
208 of the stimulus after the fixation cue goes off," while the Dly Go task can be instructed as "Re-  
209 member the direction of the stimulus, then saccade to that direction after the fixation cue goes  
210 off." Therefore, the Dly Go task can be expressed as a composition of the Go task with a partic-  
211 ular working memory process. Similarly, the Anti task can be combined with the same working  
212 memory process to form the Dly Anti task.

213 Here we test whether the network developed compositional representations for tasks, even  
214 when it was never explicitly provided with the relationships between tasks. For the sake of  
215 simplicity, we studied the representation of each task as a single high-dimensional vector. To  
216 compute this "task vector", we averaged neural activities across all possible stimulus conditions  
217 within each task and focused on the steady-state response during the stimulus epoch (**Fig. 5a**).  
218 Most tasks studied here begin with a stimulus epoch, so the neural population state near the end  
219 of stimulus presentation is potentially representative of how the network processed the stimulus  
220 in a particular task to meet the computational need of subsequent behavioral epochs. Indeed,  
221 this idea is confirmed using principal component analysis, which revealed that task vectors in  
222 the state space spanned by the top two principal components are distinct for all twenty tasks  
223 (**Supplementary Fig. 5**).

224 When plotting the task vectors representing the Go, Dly Go, Anti, and Dly Anti tasks, we found  
225 that the vector pointing from the Go vector towards the Dly Go vector is very similar to the vector  
226 pointing from the Anti vector to the Dly Anti vector (**Fig. 5b**). This finding is surprisingly robust  
227 and becomes even more apparent when we combined results from many networks (**Fig. 5c**). The



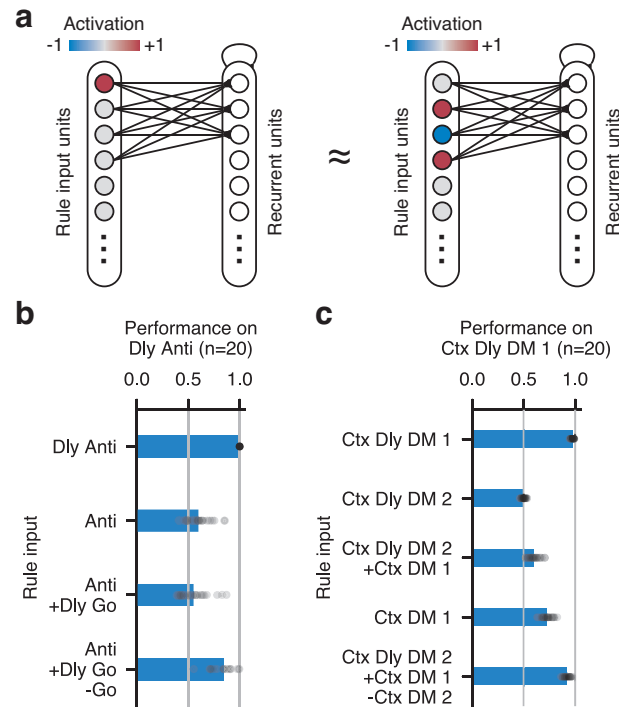
**Figure 5.** Compositional representation of tasks in state space. **(a)** The representation of each task is the population activity of the recurrent network at the end of the stimulus presentation, averaged across different stimulus conditions. **(b)** Representations of the Go, Dly Go, Anti, Dly Anti tasks in the space spanned by the top two principal components (PCs) for a sample network. For better comparison across networks, the top two PCs are rotated and reflected (rPCs) to form the two axes (see Online Methods). **(c)** The same analysis as in **(b)** is performed for 20 networks, and the results are overlaid. **(d)** Representations of the Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2 tasks in the top two PCs for a sample network. **(e)** The same analysis as in **(d)** for 20 networks.

228 Go-to-Dly Go vector and the Anti-to-Dly Anti vector presumably reflect the cognitive process  
229 of working memory. Similar findings are made with another set of tasks. The vector pointing  
230 from the Ctx DM 1 task to the Ctx DM 2 task is similar to the vector pointing from the Ctx Dly  
231 DM 1 task to the Ctx Dly DM 2 task (**Fig. 5d,e**). The Ctx DM 1-to-Ctx DM 2 vector reflects the  
232 difference between the gating modality 1 and the gating modality 2 processes. These results  
233 suggest that sensory, cognitive, and motor processes can be represented as vectors in the task  
234 space. Therefore, the representation of a task can potentially be expressed as a linear summation  
235 of vectors representing the underlying sensory, cognitive, and motor processes. This finding is  
236 reminiscent of previous work showing that neural networks can represent words and phrases  
237 compositionally<sup>31</sup>.

### 238 **Performing tasks with composition of rule inputs**

239 We showed that the representation of tasks could be compositional in principle. However, it is  
240 unclear whether in our network this principle of compositionality can be extended from rep-  
241 resenting to performing tasks. The network is normally instructed which task to perform by  
242 activation of the corresponding rule input unit. What would the network do in response to a  
243 compositional rule signal as a combination of several activated and deactivated rule units? We  
244 tested whether the network can perform tasks by receiving composite rule inputs (**Fig. 6a**).

245 Consider the same two sets of tasks as in **Fig. 5**. The network can perform the Dly Anti task  
246 well when provided with the particular combination of rule inputs: Anti + (Dly Go - Go) (**Fig.**  
247 **6b**). In contrast, the network fails to perform the Dly Anti task when provided with several other  
248 combinations of rule inputs (**Fig. 6b**). Similarly, the network can perform the Ctx Dly DM 1 task  
249 best when provided the composite rule inputs of Ctx Dly DM 2 + (Ctx DM 1 - Ctx DM 2) (**Fig. 6c**).  
250 In accordance with these results, we found that connection weights from individual rule input  
251 units to recurrent units also display a compositional structure (**Supplementary Fig. 6**). Together,  
252 these results further confirmed that our network learned the implicit compositional relationship  
253 between tasks. In such a network, learning a new task may not require any modification to the  
254 recurrent connections. Instead, it only requires learning the appropriate combination of rule  
255 inputs that control the information flow within the network<sup>2</sup>.



**Figure 6.** Performing tasks with algebraically composite rule inputs. **(a)** During training, a task is always instructed by activation of the corresponding rule input unit (left). After training, the network can potentially perform a task by activation or deactivation of a set of rule input units meant for other tasks (right). **(b)** The network can perform the Dly Anti task well if given the Dly Anti rule input or the Anti + (Dly Go - Go) rule input. The network fails to perform the Dly Anti task when provided other combinations of rule inputs. **(c)** Similarly, the network can perform the Ctx Dly DM 1 task well when provided with the Ctx Dly DM 2 + (Ctx DM 1 - Ctx DM 2) rule input. Circles represent results of individual networks, while bars represent median performances of 20 networks.



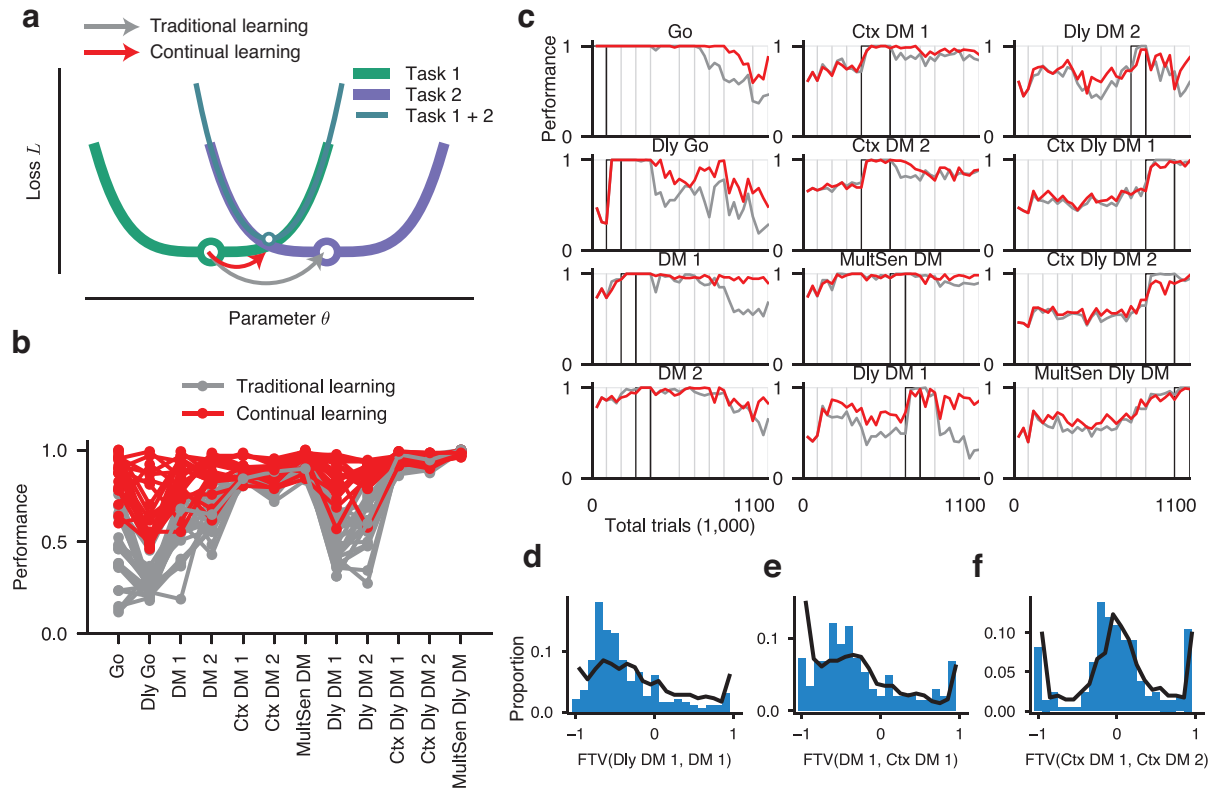
## 256 **Continual training of many cognitive tasks**

257 In humans and other animals, the performance of a well-trained task can be retained, even with-  
258 out re-training, for months or even years. However, when using traditional network training  
259 techniques, artificial neural networks rapidly forget previously learned tasks after being exposed  
260 to new tasks. This failure of retaining memories during sequential training of tasks, termed  
261 "catastrophic forgetting," is inevitable when using common network architectures and training  
262 methods<sup>32,33</sup>. Network parameters (such as connection weights) optimal for a new task can be  
263 destructive for old tasks (**Fig. 7a**). Recent work proposed several continual learning methods to  
264 battle catastrophic forgetting<sup>32-34</sup>. These methods typically involve selective protection of con-  
265 nection weights that are deemed important for previously learned tasks.

266 By employing one such technique<sup>33</sup>, we were able to substantially improve the performance  
267 of networks that are sequentially trained on a set of cognitive tasks (**Fig. 7b**). The continual learn-  
268 ing technique is especially effective at helping the network retain performance of tasks learned  
269 earlier. For example, the network can retain high performance in a working memory task after  
270 successfully learning ten additional tasks (**Fig. 7c**). We analyzed the FTV distributions for three  
271 example pairs of tasks in the continual learning networks (**Fig. 7d-f**). The shapes of these FTV  
272 distributions can be markedly different from the corresponding ones of the interleaved-training  
273 networks (**Fig. 7d,e, Fig. 4b,c**). It is possible that this result depends on factors in the continual  
274 learning, such as the order of individual tasks used during training, more careful comparisons  
275 are needed in future studies. Nevertheless, our findings suggest that sequential training of tasks  
276 could drastically shape neural network representations.

## 277 **DISCUSSION**

278 Higher-order cortical areas, especially the lateral prefrontal cortex, are remarkably versatile for  
279 their engagement in a wide gamut of cognitive functions. Here we investigated how multiple  
280 cognitive tasks are represented in a single recurrent neural network model. First, we demon-  
281 strated how the trained neural network could be dissected and understood for a family of tasks.  
282 Next, we identified clusters of units that are each specialized for a subset of tasks. Each cluster  
283 potentially represents a particular sequence of the sensori-motor events and a subset of cogni-



**Figure 7.** Sequential training of cognitive tasks. **(a)** Schematics of continual learning. The network learns to perform a task by modifying parameters to minimize the loss function for this task. When a network is trained on two tasks sequentially with traditional learning techniques (gray arrow), training for the second task can easily result in the failure of performing the first task, because the minima (circle) of the loss functions of tasks 1 (green) and 2 (purple) are far apart. Continual learning techniques can protect previously-learned tasks by preventing large changes of important network parameters (red arrow). Arrows show changes of an example parameter  $\theta$  when task 2 is trained after task 1 is already learned. **(b)** Final performance across all trained tasks with traditional (gray) or continual (red) learning techniques. Only 12 tasks are trained due to difficulty of learning more tasks even with continual learning techniques. Lines represent results of individual networks. **(c)** Performance of all tasks during sequential training of one network with traditional (gray) or continual (red) learning techniques. For each task, the black box indicates the period in which this task is trained. **(d-f)** Fractional variance distributions for three pairs of tasks. Blue: distribution for one sample network. Black: averaged distribution over 20 networks.

284 tive (such as working memory, categorization, decision-making, inhibitory control) processes  
285 that are the building blocks for flexible behavior. We observed a close match between the task  
286 selectivity of each cluster and its causal role, which suggests that analysis of neural activity can  
287 provide meaningful functional insights in trained neural networks. We proposed a measure, the  
288 fractional task variance, that probes the neural relationship between a pair of tasks at the single-  
289 neuron level. This measure allowed us to summarize five distinct and typical kinds of neural  
290 relationships in our network. This measure can be readily applied to firing activity of single units  
291 recorded from animals performing two or more tasks. Surprisingly, we found that the represen-  
292 tation of tasks in our network is compositional, a critical feature for cognitive flexibility. By virtue  
293 of the compositionality, a task can be correctly instructed by composing instructions for other  
294 tasks. Finally, using a recently proposed continual learning technique, we can train the network  
295 to learn many tasks sequentially.

296 Monkeys, and in some cases rodents, can be trained to alternate between two tasks<sup>11,22,25,35,36</sup>.  
297 Single-unit recordings from these experiments can potentially be analyzed to compute the frac-  
298 tion task variance distributions. Theoretical studies argued that for maximum cognitive flex-  
299 ibility, prefrontal neurons should be selective to mixtures of multiple task variables<sup>37</sup>. Mixed  
300 selectivity neurons are indeed ubiquitous within the prefrontal cortex<sup>10</sup>. We showed that most  
301 units in our network are strongly selective to rules (**Fig. 3**). Meanwhile, these units are selective  
302 to other aspects of tasks (otherwise their task variances would be zero). For example, the Anti  
303 units (**Fig. 2**) are highly activated only during the Anti tasks and when the stimulus is in their  
304 preferred directions. Therefore, units in our network display strong nonlinear mixed selectivity,  
305 as found in neurons of the prefrontal cortex<sup>10</sup>. Conceptually, this work extends the notion of  
306 mixed selectivity from within a single task to across multiple tasks.

307 Multiple cognitive tasks are more common in human imaging studies. In a series of exper-  
308 iments, Cole and colleagues trained humans to perform 64 cognitive tasks following composi-  
309 tional rule instructions<sup>6,38</sup>. They trained linear classifiers to decode rules from prefrontal neural  
310 activity patterns. These classifiers can significantly generalize to novel tasks<sup>6</sup>, consistent with a  
311 compositional neural representation of rules. Although trained with discrete rule instructions,  
312 our network develops a clear compositional structure in its representations, as shown using the  
313 population activity at a single time point (near the end of stimulus presentation). Temporal dy-

314 namics in neural circuits are ubiquitous during cognitive tasks<sup>39</sup> and are potentially critical for  
315 cognitive computations<sup>18</sup>, so the study of steady-state responses here is merely a first step to-  
316 wards understanding the dynamical representation of tasks. Future work could study how dy-  
317 namical representations of tasks are related to one another in the state space. Cole et al. found  
318 that humans can rapidly adapt to new tasks by adjusting the functional connectivity patterns  
319 of parietal-frontal flexible hubs<sup>38</sup>. In the future, graph-theoretic analysis can be used to test  
320 whether our trained network developed flexible hubs that coordinate information flow across  
321 the network. There exists a structural hierarchy within the human prefrontal cortex, with more  
322 abstract cognitive processes being represented in the more anterior areas<sup>40,41</sup>. It is unclear if our  
323 trained network developed hierarchical representations of cognitive processes or tasks. If it did,  
324 a subset of units should represent more abstract aspects of the tasks, while other units represent  
325 the concrete, sensorimotor aspects. This question is hard to address for now because the 20 tasks  
326 we chose are not organized in a clearly hierarchical way<sup>40</sup>.

327 Training artificial neural networks for multiple tasks has a long history in the field of machine  
328 learning<sup>42</sup>. However, it has mainly been used as a method to improve training and generaliza-  
329 tion. There were few studies on the representation of task structure or task set in trained net-  
330 works. Modern artificial neural networks are capable of highly complex tasks, such as playing  
331 Atari games<sup>43</sup>, which likely involve a range of cognitive skills. However, in contrast to cognitive  
332 tasks that are specifically designed to shed light on neural mechanisms of cognition, complex  
333 real-life tasks remain challenging to analyze. In principle, we can strike a balance between the  
334 two approaches by designing a set of tasks that are complex enough, yet still amenable to analy-  
335 sis. The ability to "open the box" and elucidate the inner working of the network after training is  
336 crucial for understanding neural mechanisms of cognition in neuroscience.

337 Like other works on trained neural networks<sup>11,14-19,44</sup>, the machine learning protocol we  
338 used is not validated biologically. Besides, our RNN consists of a single neural population, in  
339 contrast to the brain system where a number of interacting brain regions are engaged in a cog-  
340 nitive task<sup>22,45</sup>. Although our neural network model developed functionally specialized clusters  
341 of units through training, it is unclear how to map them onto different brain areas. Furthermore,  
342 in our network, a rule input is explicitly provided throughout the trial, therefore there is no need  
343 for the network to hold the "task set" internally using persistent activity<sup>4,5</sup>. This, however, can

344 be remedied by providing the rule cue only at the beginning of each trial, which would encour-  
345 age the network to internally sustain the task set. We can even ask the network to figure out a  
346 task rule by trial-and-error<sup>46</sup>. In spite of these concerns, our approach offers an efficient com-  
347 putational platform to test hypotheses about neural representations and mechanisms that could  
348 guide experiments and data analysis. Furthermore, this approach can yield new conceptual in-  
349 sights, as shown here by the finding of compositional task representation. Future progress in  
350 this direction, at the interface between neuroscience and artificial intelligence, will advance our  
351 understanding of flexible behavior in many cognitive tasks.

## 352 **ACKNOWLEDGMENTS**

353 We thank current and former members of the Wang lab, especially S.Y. Li, O. Marschall, and M.  
354 Joglekar, and E. Ohran for fruitful discussions; J.A. Li, J.D. Murray, D. Ehrlich, and J. Jaramillo for  
355 critical comments on the manuscript; and S. Wang for assistance with the NYU HPC clusters.  
356 This work was supported by an Office of Naval Research Grant N00014-13-1-0297, a National  
357 Science Foundation Grant Number 16-31586, and a Google Computational Neuroscience Grant  
358 (X.J.W.) and a Samuel J. and Joan B. Williamson Fellowship (G.R.Y.).

## 359 **AUTHOR CONTRIBUTIONS**

360 G.R.Y. and X.J.W. designed the study. G.R.Y., H.F.S, W.T.N, and X.J.W. had frequent discussions.  
361 G.R.Y. performed the research. G.R.Y., H.F.S, W.T.N, and X.J.W. wrote the manuscript.

## 362 **COMPETING FINANCIAL INTERESTS**

363 The authors declare no competing financial interests.

## 364 **DATA AVAILABILITY STATEMENT**

365 All codes will be available at publication.

## 366 ONLINE METHODS

### 367 Code availability

368 All codes are available on GitHub.

### 369 Network structure

370 The recurrent neural networks shown in the main text all contain  $N_{\text{rec}} = 256$  units. The results  
371 are largely insensitive to the network size. Similar results were obtained in networks of sizes  
372 between 128 and 512 units (the range we tested). The network is a time-discretized recurrent  
373 neural network with positive activity<sup>15</sup>. Before time-discretization, the network activity  $\mathbf{r}$  follows  
374 a continuous dynamical equation

$$\tau \frac{d\mathbf{r}}{dt} = -\mathbf{r} + f(W^{\text{rec}}\mathbf{r} + W^{\text{in}}\mathbf{u} + \mathbf{b}) + \sqrt{2\tau\sigma_{\text{rec}}^2}\xi. \quad (2)$$

375 In this equation,  $\mathbf{u}$  is the input to the network,  $\mathbf{b}$  is the bias or background input,  $\tau = 100\text{ms}$  is  
376 the neuronal time constant,  $f(\cdot)$  is the neuronal nonlinearity that keeps the unit activity non-  
377 negative,  $\xi$  are  $N_{\text{rec}}$  independent Gaussian white noise processes with zero mean and unit vari-  
378 ance, and  $\sigma_{\text{rec}} = 0.05$  is the strength of the noise. In particular, we use a standard softplus func-  
379 tion

$$f(x) = \log(1 + \exp(x)), \quad (3)$$

380 which after re-parameterization is very similar to a neuronal nonlinearity, i.e., f-I curve, com-  
381 monly used in previous neural circuit modelings<sup>28</sup>. A set of output units  $\mathbf{z}$  read out nonlinearly  
382 from the network,

$$\mathbf{z} = g(W^{\text{out}}\mathbf{r}), \quad (4)$$

383 where  $g(x) = 1/(1 + \exp(-x))$  is the logistic function, bounding output activities between 0 and  
384 1.  $W^{\text{in}}$ ,  $W^{\text{rec}}$ ,  $W^{\text{out}}$  are the input, recurrent, and output connection matrices respectively.

385 After using the first-order Euler approximation with a time discretization step  $\Delta t$ , we have

$$\mathbf{r}_t = (1 - \alpha)\mathbf{r}_{t-1} + \alpha \cdot f(W^{\text{rec}}\mathbf{r}_{t-1} + W^{\text{in}}\mathbf{u}_t + \mathbf{b}) + \sqrt{2\alpha\sigma_{\text{rec}}^2}\mathbf{N}(0, 1). \quad (5)$$

386 Here  $\alpha \equiv \Delta t/\tau$ , and  $\mathbf{N}(0, 1)$  stands for the standard normal distribution. We use a discretization  
387 step  $\Delta t = 20\text{ms}$ . We imposed no constraint on the sign or the structure of the weight matrices  
388  $W^{\text{in}}, W^{\text{rec}}, W^{\text{out}}$ . The network and the training are implemented in TensorFlow<sup>47</sup>.

389 The network receives four types of noisy inputs,

$$\mathbf{u} = (u_{\text{fix}}, \mathbf{u}_{\text{mod1}}, \mathbf{u}_{\text{mod2}}, \mathbf{u}_{\text{rule}}) + \mathbf{u}_{\text{noise}}. \quad (6)$$

$$\mathbf{u}_{\text{noise}} = \sqrt{\frac{2}{\alpha}} \sigma_{\text{in}} \mathbf{N}(0, 1). \quad (7)$$

390 Here the input noise strength  $\mathbf{N}(0, 1) = 0.01$ . The fixation input  $u_{\text{fix}}$  is typically at the high value of  
391 1 when the network should fixate. The fixation input goes to zero when the network is required to  
392 respond. The stimulus inputs  $\mathbf{u}_{\text{mod1}}$  and  $\mathbf{u}_{\text{mod2}}$  comprise two "rings" of units, each representing  
393 a one-dimensional circular variable described by the degree around a circle. Each ring contains  
394 32 units, whose preferred directions are uniformly spaced from 0 to  $2\pi$ . For unit  $i$  with a preferred  
395 direction  $\theta_i$ , its activity for a stimulus presented at direction  $\psi$  is

$$u_i = \gamma \cdot 0.8 \exp \left[ -\frac{1}{2} \left( \frac{8|\psi - \theta_i|}{\pi} \right)^2 \right], \quad (8)$$

396 where  $\gamma$  is the strength of the stimulus. For multiple stimuli, input activities are added together.  
397 The network also receives a set of rule inputs  $\mathbf{u}_{\text{rule}}$  that encode which task the network is sup-  
398 posed to perform on each trial. Normally,  $\mathbf{u}_{\text{rule}}$  is a one-hot vector. That means the rule input  
399 unit corresponding to the current task is activated at 1, while other rule input units remain at  
400 0. Therefore the number of rule input units equals to the number of tasks trained. For compo-  
401 sitional rule inputs (**Fig. 6**), the activation of rule input units can be an arbitrary pattern. For  
402 example, for the combined rule input Anti + (Dly Go - Go), the activities of the rule input units  
403 corresponding to the Go, Dly Go, and Anti tasks are -1, +1, and +1 respectively. In total there are  
404  $N_{\text{in}} = 1 + 32 \times 2 + 20 = 85$  input units.

405 The network projects to an output ring  $\mathbf{z}_{\text{out}}$ , which also contains 32 units. The output ring  
406 units encode the response directions using similar tuning curves to the ones used for the input  
407 rings. In addition, the network projects to a fixation output unit  $\mathbf{z}_{\text{fix}}$ , which should be at the high  
408 activity value of 1 before the response and at 0 once a response is generated. In total there are  
409  $N_{\text{out}} = 1 + 32 = 33$  output units.

410 We lesion a network unit by setting to zero its projection weights to all recurrent and output  
411 units.

## 412 **Tasks and performances**

413 Here we first describe the common setup for the 20 tasks trained. Deviations from the common  
414 setup will be described below individually. The rule input unit corresponding to the current  
415 task will be activated throughout the whole trial. The network receives a fixation input, which  
416 is activated from the beginning of the trial. When the fixation input is on, the network should  
417 fixate by having the fixation output unit at a high activity  $\hat{z}_{\text{fix}} = 0.85$ . The offset of the fixation  
418 input usually indicates the onset of the response or go epoch, when the network needs to report  
419 the response direction through activities of the output ring. During the response epoch, the  
420 fixation output unit has a target output of  $\hat{z}_{\text{fix}} = 0.05$ . For a target response direction  $\psi$ , the target  
421 output activity of an output unit  $i$  is

$$\hat{z}_i = 0.8 \exp \left[ -\frac{1}{2} \left( \frac{8|\psi - \psi_i|}{\pi} \right)^2 \right] + 0.05, \quad (9)$$

422 where  $\psi_i$  is the preferred response direction of unit  $i$ . When no response is required, the target  
423 output activity is fixed at  $\hat{z}_i = 0.05$ . The network also receives one or two stimuli. Each stimulus  
424 contains information from modality 1, 2, or both. When there is only one stimulus, the direction  
425 of the stimulus is drawn from a uniform distribution between 0 and 360 degree.

426 A trial is considered correct only if the network correctly maintained fixation and responded  
427 to the correct direction. The response direction of the network is read out using a population  
428 vector method. The decoded response direction is considered correct if it is within 36 degrees of  
429 the target direction. If the activity of the fixation output falls below 0.5, the network is considered  
430 to have broken fixation.

431 The discrimination thresholds  $a$  in **Supplementary Fig. 2** are obtained by fitting Weibull  
432 functions to performances  $p$  as a function of coherences  $c$  at a fixed stimulus duration,

$$p = 1 - 0.5 \exp(-(c/a)^b). \quad (10)$$

433 Each task can be separated into distinct epochs. Fixation (fix) epoch is the period before  
434 any stimulus is shown. It is followed by the stimulus epoch 1 (stim1). If there are two stimuli  
435 separated in time, then the period between the two stimuli is the delay epoch, and the second  
436 stimulus is shown in the stimulus epoch 2 (stim2). The period when the network should re-  
437 spond is the go epoch. The duration of the fixation, stim1, delay1, stim2, and go epochs are



438  $T_{\text{fix}}, T_{\text{stim1}}, T_{\text{delay1}}, T_{\text{stim2}}, T_{\text{go}}$  respectively. For convenience, we grouped the 20 tasks into five task  
439 families: the Go, Anti, Decision-Making (DM), Delayed Decision-Making (Dly DM), and Match-  
440 ing families.

441 **Go task family.** This family of tasks includes the Go, RT Go, and Dly Go tasks. In all three tasks, a  
442 single stimulus is randomly shown in either modality 1 or 2, and the response should be made in  
443 the direction of the stimulus. These three tasks differ in their stimulus onset and offset times. In  
444 the Go task, the stimulus appears before the fixation cue goes off. In the RT Go task, the fixation  
445 input never goes off, and the network should respond as soon as the stimulus appears. In the Dly  
446 Go task, a stimulus appears briefly and is followed by a delay period until the fixation cue goes  
447 off. The Dly Go task is similar to the memory-guided saccade task<sup>20</sup>.

448 For the Go task,

$$T_{\text{stim1}} \sim \mathbf{U}(500, 1500). \quad (11)$$

449  $\mathbf{U}(t_1, t_2)$  is a uniform distribution between  $t_1$  and  $t_2$ . The unit for time is ms and is omitted for  
450 brevity. For the RT Go task,

$$T_{\text{stim1}} \sim \mathbf{U}(500, 2500). \quad (12)$$

451 For the Dly Go tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}). \quad (13)$$

452 Here  $\mathbf{U}(\{a_1, \dots, a_n\})$  denotes a discrete uniform distribution over the set  $\{a_1, \dots, a_n\}$ .

453 **Anti task family.** This family includes the Anti, RT Anti, and Dly Anti tasks. These three tasks  
454 are the same as their counterpart Go-family tasks, except that the response should be made to  
455 the opposite direction of the stimulus.

456 **DM family.** This family includes five perceptual decision making tasks: the DM 1, DM 2, Ctx  
457 DM 1, Ctx DM 2, and MultSen DM tasks. In each trial, two stimuli are shown simultaneously and  
458 are presented till the end of the trial. Stimulus 1 is drawn randomly between 0 and 360 degree,  
459 while stimulus 2 is drawn uniformly between 90 and 270 degree away from stimulus 1. In DM 1,

460 the two stimuli only appear in modality 1, while in DM 2, the two stimuli only appear in modality  
461 2. In DM 1 and DM 2, the correct response should be made to the direction of the stronger  
462 stimulus (the stimulus with higher  $\gamma$ ). In Ctx DM 1, Ctx DM 2, and MultSen DM tasks, each  
463 stimulus appears in both modality 1 and 2. In the Ctx DM 1 task, information from modality 2  
464 should be ignored, and the correct response should be made to the stronger stimulus in modality  
465 1. In the Ctx DM 2 task, information from modality 1 should be ignored. In the MultSen DM task,  
466 the correct response should be made to the stimulus that has a stronger combined strength in  
467 modalities 1 and 2.

468 The DM 1 and DM 2 tasks are inspired from classical perceptual decision making tasks based  
469 on random-dot motion stimuli<sup>21</sup>. In random-dot motion tasks, there is only one stimulus, the  
470 coherence of which is varied across trials. Following the tradition of Wang, 2002<sup>30</sup>, we use two  
471 input stimuli to model momentary motion evidence towards the two target directions. When  
472 the two stimuli have the same strengths ( $\gamma_1 = \gamma_2$ ), there is no net evidence towards any target  
473 direction, mimicking the condition of 0 motion coherence in the random-dot motion task. A  
474 stronger difference in the stimulus strengths emulates a stronger motion coherence. For a co-  
475 herence  $c$  representing net evidence for the direction of stimulus 1, the strengths of stimulus 1  
476 and 2 ( $\gamma_1, \gamma_2$ ) are set as

$$\gamma_{1,modi} = \bar{\gamma} + c, \quad \gamma_{2,modi} = \bar{\gamma} - c, \quad (14)$$

477 respectively, where  $i \in 1, 2$  is the modality. Here  $\bar{\gamma}$  is the average strength of the two stimuli. For  
478 each trial, we draw  $\bar{\gamma}$  from a uniform distribution around 1,  $\bar{\gamma} \sim \mathbf{U}(0.8, 1.2)$ . Indeed, in all DM-  
479 family tasks and Dly DM-family tasks, there is a single coherence  $c$  in each trial that determines  
480 the overall strength of net evidence towards the direction represented by stimulus 1. For all DM  
481 family tasks,

$$c \sim \mathbf{U}(\{-0.08, -0.04, -0.02, -0.01, 0.01, 0.02, 0.04, 0.08\}). \quad (15)$$

482 The duration of stimulus 1, which is fixed in each trial, is drawn from the following distribu-  
483 tion,

$$T_{stim1} \sim \mathbf{U}(\{400, 800, 1600\}). \quad (16)$$

484 Indeed, all tasks from the DM family use the same distribution for  $T_{stim1}$ . And since the two  
485 stimuli are shown simultaneously,  $T_{stim1} = T_{stim2}$ .

486 The Ctx DM 1 and Ctx DM 2 tasks are inspired from context-dependent decision-making  
487 tasks performed by macaque monkeys<sup>11</sup>. Now each stimulus is presented in both modalities at  
488 the same direction, with strengths  $\gamma_{1,\text{mod1}}, \gamma_{1,\text{mod2}}$  for stimulus 1, and  $\gamma_{2,\text{mod1}}, \gamma_{2,\text{mod2}}$  for stimu-  
489 lus 2. The stimulus strengths are determined by the coherence for modality 1 and 2 ( $c_{\text{mod1}}, c_{\text{mod2}}$ ),  
490 so we have

$$\gamma_{1,\text{mod1}} = \bar{\gamma}_{\text{mod1}} + c_{\text{mod1}}, \gamma_{2,\text{mod1}} = \bar{\gamma}_{\text{mod1}} - c_{\text{mod1}}. \quad (17)$$

491 Similar equation holds for modality 2 as well.  $c_{\text{mod1}}$  and  $c_{\text{mod2}}$  are drawn independently from the  
492 same distribution. In Ctx DM 1,  $c = c_{\text{mod1}}$ , while in Ctx DM 2,  $c = c_{\text{mod2}}$ .  $\bar{\gamma}_{\text{mod1}}$  and  $\bar{\gamma}_{\text{mod2}}$  are also  
493 drawn from  $\mathbf{U}(0.8, 1.2)$ . In the original Mante task<sup>11</sup>, there is an additional delay period between  
494 the stimuli and the response period, which is not included here.

495 The MultSen DM task mimics a multi-sensory integration task<sup>23</sup>. The setup of stimulus is  
496 similar to those in the Ctx DM 1 and Ctx DM 2 tasks, except that the network should integrate  
497 information from both modalities and the stronger stimulus is the one with higher averaged  
498 strength from modality 1 and 2. The overall coherence  $c = (c_{\text{mod1}} + c_{\text{mod2}})/2$ . We determine all  
499 four strengths with the following procedure. First we determine the average strength of stimulus  
500 1 across both modalities,  $\gamma_1$ , and the average strength of stimulus 2,  $\gamma_2$ .

$$\gamma_1 = \bar{\gamma} + c, \gamma_2 = \bar{\gamma} - c. \quad (18)$$

501 Here  $\bar{\gamma}$  and  $c$  both follow the same distributions as other DM-family tasks. Then we set

$$\gamma_{1,\text{mod1}} = \gamma_1(1 + \Delta_1), \gamma_{1,\text{mod2}} = \gamma_1(1 - \Delta_1), \quad (19)$$

502 where  $\Delta_1 \sim \mathbf{U}(0.1, 0.4) \cup \mathbf{U}(-0.4, -0.1)$ . Similarly for stimulus 2.

503 **Dly DM family.** This family includes Dly DM 1, Dly DM 2, Ctx Dly DM 1, Ctx Dly DM 2. These  
504 tasks are similar to the corresponding tasks in the DM family, except that in the Dly DM family  
505 tasks, the two stimuli are separated in time. The Dly DM 1 and Dly DM 2 tasks are inspired by  
506 the classical parametric working memory task developed by Romo and colleagues<sup>24</sup>. The two  
507 stimuli are both shown briefly and are separated by a delay period. Another short delay period  
508 follows the offset of the second stimulus.

509 For all Dly DM family tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}), c \sim \mathbf{U}(\{-0.32, -0.16, -0.08, 0.08, 0.16, 0.32\}). \quad (20)$$

510 And,  $T_{\text{stim1}} = T_{\text{stim2}} = 300$ .

511 **Matching family.** This family of tasks includes the DMS, DNMS, DMC, DNMC tasks. In these  
512 tasks, two stimuli are presented consecutively and separated by a delay period. Each stimulus  
513 can appear in either modality 1 or 2. The network response depends on whether or not the  
514 two stimuli are "matched." In the DMS and DNMS tasks, two stimuli are matched if they point  
515 towards the same direction, regardless of their modalities. In DMC and DNMC tasks, two stimuli  
516 are matched if their directions belong to the same category. The first category ranges from 0 to  
517 180 degrees, while the rest from 180 to 360 degrees belongs to the second category. In the DMS  
518 and DMC tasks, the network should respond towards the direction of the second stimulus if the  
519 two stimuli are matched and maintain fixation otherwise. In the DNMS and DNMC tasks, the  
520 network should respond only if the two stimuli are not matched, i.e., a non-match, and fixate  
521 when it is a match.

522 During training of these tasks, half of the trials are matching, and the other half are non-  
523 matching. In DMS and DNMS tasks, stimulus 1 is always drawn randomly. In half of the trials,  
524 stimulus 2 appears at the same direction as stimulus 1. In the other half, stimulus 2 is drawn  
525 randomly between 10 and 350 degree away from stimulus 1. In DMC and DNMC tasks, both  
526 stimulus 1 and 2 are drawn randomly and independently from the uniform distribution

$$\mathbf{U}(\{18, 54, 90, 126, 162, 198, 234, 270, 306, 342\}). \quad (21)$$

527 In all Matching family tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}). \quad (22)$$

528 Also, match trials and non-match trials always appear with equal probability.

## 529 **Training procedure**

530 The loss  $\mathcal{L}$  to be minimized is computed by time-averaging the squared errors between the net-  
531 work output  $\mathbf{z}(t)$  and the target output  $\hat{\mathbf{z}}(t)$ .

$$\mathcal{L} = \mathcal{L}_{\text{mse}} \equiv \langle \mathbf{m}_{i,t} (\mathbf{z}_{i,t} - \hat{\mathbf{z}}_{i,t})^2 \rangle_{i,t}. \quad (23)$$

532 Here  $i$  is the index of the output units. The squared errors at different time points and of different  
533 output units are potentially weighted differently according to the non-negative mask matrix  $\mathbf{m}_{i,t}$ .  
534 For the output ring units, before the response epoch, we have  $\mathbf{m}_{i,t} = 1$ . The first 100ms of the  
535 response epoch is a grace period with  $\mathbf{m}_{i,t} = 0$ , while for the rest of the response epoch,  $\mathbf{m}_{i,t} = 5$ .  
536 For the fixation output unit,  $\mathbf{m}_{i,t}$  is two times stronger than the mask for the output ring units.

537 The training is performed with Adam, a powerful variant of stochastic gradient descent<sup>48</sup>. We  
538 used the default set of parameters. The learning rate is 0.001, the decay rate for the 1st and 2nd  
539 moment estimates are 0.9 and 0.999 respectively.

540 The recurrent connection matrix is initialized with a scaled identity matrix  $q \cdot \mathbf{1}^{49}$ , where  $\mathbf{1}$   
541 is the identity matrix. We chose  $q = 0.54$  such that the gradient is roughly preserved during  
542 backpropagation when the network is initialized. The input and output connection weights  
543 are initialized as independent Gaussian random variables with mean 0, and standard deviations  
544  $1/\sqrt{N_{\text{in}}}$  and  $0.4/\sqrt{N_{\text{rec}}}$  respectively. The standard deviation value for the output weights is cho-  
545 sen to prevent saturation of output units after initialization.

546 During training, we randomly interleaved all the tasks with equal probabilities, except for the  
547 Ctx DM 1 and Ctx DM 2 tasks that appear five times more frequently, because without sufficient  
548 training, the network gets stuck at an alternative strategy. Instead of correctly ignoring modality  
549 1 or 2, the network can choose to ignore the context and integrate information from both modal-  
550 ities equally. This strategy gives the network an accuracy close to 75%. During training, we used  
551 mini-batches of 64 trials, in which all trials are generated from the same task for computational  
552 efficiency.

### 553 **Analysis of the Anti task family**

554 Anti units in **Fig. 2** are defined as those units that have higher summed task variance (see next  
555 section for definition) for the Anti family of tasks ( $S_{\text{Anti}} = \text{Anti, RT Anti, Dly Anti}$ ) than for all other  
556 tasks. So a unit  $i$  is an Anti unit if

$$\sum_{A \in S_{\text{Anti}}} \text{TV}_i(A) > \sum_{A \notin S_{\text{Anti}}} \text{TV}_i(A). \quad (24)$$

## 557 Task variance analysis

558 A central goal of our analysis was to determine whether individual units within the network are  
559 selective to different tasks, or whether units tended to be similarly selective to all tasks. To quan-  
560 tify how selective a unit is in one task, we defined a task variance metric. To compute the task  
561 variance  $TV_i(A)$  for task A and unit  $i$ , we ran the network for many stimulus conditions that span  
562 the space of possible stimuli. For example, in the DM family tasks, we ran the network for stimuli  
563 with directions ranging from 0 to 360 degrees and with coherences ranging from almost 0 to 0.2.  
564 After running the network for many stimulus conditions, we computed the variance across stim-  
565 ulus conditions (trials) at each time point for a specific unit then averaged the variance across all  
566 time points to get the final task variance for this unit. The fixation epoch is excluded from this  
567 analysis. This process was repeated for each unit in the network. Therefore

$$TV_i(A) = \langle [r_i(j, t) - \langle r_i(j, t) \rangle_j]^2 \rangle_{j,t}, \quad (25)$$

568 where  $r_i(j, t)$  is the activity of unit  $i$  on time  $t$  of trial  $j$ . In **Fig. 2,3,4**, we only analyzed active  
569 units, defined as those that have summed task variance across tasks higher than a threshold,  
570  $10^{-3}$ . The results do not depend strongly on the choice of the threshold. This procedure prevents  
571 units with extremely low task variance from being included in the analysis.

572 By computing each unit's selectivity across different stimulus conditions, we naturally in-  
573 clude the selectivity to motor outputs, because motor outputs depend ultimately on the stimuli.  
574 A unit that is only selective to motor outputs or other cognitive variables in a task will still have  
575 a non-zero task variance. Units that are purely selective to rules and/or time will, however, have  
576 zero task variance and therefore be excluded from our analysis.

577 The clustering of units based on their task variance patterns in **Fig. 3** uses K-means cluster-  
578 ing from the Python package scikit-learn. To assess how well a clustering configuration is, we  
579 computed its silhouette coefficient based on intra-cluster and inter-cluster distances. A higher  
580 silhouette coefficient means a better clustering. The optimal number of clusters  $\tilde{k}$  is determined  
581 by choosing the first  $k$  such that the silhouette coefficient for  $k + 1$  clusters is worse than  $k$  clus-  
582 ters.

583 In **Fig. 3d**, we visualize the clustering using t-distributed Stochastic Neighbor Embedding  
584 (tSNE). For each unit, the normalized task variances across all tasks form a 20 dimensional vec-

585 tor that is then embedded in a 2-dimensional space. For the tSNE method, we used the exact  
586 method for gradient calculation, a learning rate of 100, and a perplexity of 30.

587 The fractional task variance with respect to tasks A and B is

$$\text{FTV}_i(A, B) = \frac{\text{TV}_i(A) - \text{TV}_i(B)}{\text{TV}_i(A) + \text{TV}_i(B)}. \quad (26)$$

588 To obtain a statistical baseline for the FTV distributions as in **Supplementary Fig. 4**, we trans-  
589 form the neural activities of the network with a random orthogonal matrix before computing the  
590 task variance. For each network, we generate a random orthogonal matrix  $M$  using the Python  
591 package Scipy. All network activities are multiplied by this matrix  $M$  to obtain a rotated version  
592 of the original neural representation.

$$\mathbf{r}_t^{\text{rot}} = M\mathbf{r}_t. \quad (27)$$

593 Since multiplying neural activities by an orthogonal matrix is equivalent to rotating the neural  
594 representation in state space, this procedure will preserve results from state space analysis. We  
595 then compute task variances and fractional task variances using the rotated neural activities. The  
596 FTV distributions using the rotated activities are clearly different from the original FTV distribu-  
597 tions.

### 598 **State-space analysis**

599 To compute the representation of a task in the state space, we first computed the neural activ-  
600 ities across all possible stimulus conditions, then we averaged across all these conditions. For  
601 simplicity of the analysis, we chose to analyze only the steady state responses during the stimu-  
602 lus epoch. We do so by focusing on the last time point of the stimulus epoch,  $t_{\text{stim1, end}}$ . So the  
603 representation of task A is

$$\tilde{\mathbf{r}} = \langle \mathbf{r}(j, t_{\text{stim1, end}}) \rangle_j, \quad (28)$$

604 where  $\mathbf{r}(j, t)$  is the vector of network activities at trial  $j$  and time  $t$  during task A.

605 For each set of tasks, we performed principal component analysis to get the lower dimen-  
606 sional representation. We repeated this process for different networks. The representations of  
607 Go, Anti, Dly Go, and Dly Anti tasks are close to four vertices of a square. As a result, the top

608 two principal components have similar eigenvalues and are therefore interchangeable. To better  
609 compare across networks in **Fig. 5b,c**, we allowed a rotation and a reflection within the space  
610 spanned by the top two PCs. For each network, the rotated and reflected PCs (rPCs) are chosen  
611 such that the Go task representation lies on the positive part of the x-axis, and the Dly Go task  
612 lies below the x-axis. The representation of Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM  
613 2 tasks do not form a square, so we only allowed reflections such that Ctx Dly DM 1 is in the first  
614 quadrant. The reflected PCs are still PCs.

### 615 **Continual learning**

616 For continual learning in **Fig. 7**, tasks appear sequentially. Each task is trained for 150,000 trials.  
617 Ctx DM 1 and Ctx DM 2 are still trained together and interleaved, and so are Ctx Dly DM 1 and  
618 Ctx Dly DM 2. We added a regularizer that protects old tasks by setting an additional penalty for  
619 deviations of important synaptic weights (or other parameters)<sup>33</sup>. When training the  $\mu$ -th task,  
620 the regularizer is

$$\mathcal{L}_{\text{cont}} = c_{\text{cont}} \sum_k \Omega_k^\mu (\theta_k - \tilde{\theta}_k)^2. \quad (29)$$

621 Here  $c_{\text{cont}}$  is the overall strength of the regularizer,  $\theta_k$  denotes the  $k$ -th parameter of the network.  
622 The value of the anchor parameter  $\tilde{\theta}_k$  is the value of  $\theta_k$  at the end of the last task (the  $(\mu - 1)$ -th  
623 task). No regularizer is used when training the first task. And  $\Omega_k^\mu$  measures how important the  
624 parameter is. Notice that two recent proposals<sup>32,33</sup> for continual learning both use regularizers  
625 of this form. The two proposals differ only in how the synaptic importances are computed. We  
626 chose the method of Zenke et al. 2017, because the method of Kirkpatrick et al. 2017 measures  
627 the synaptic importance locally in the parameter space, resulting in underestimated and inac-  
628 curate synaptic importance values for our settings. In Zenke et al. 2017, the importance of one  
629 parameter is determined using this parameter's historic contribution to the change in the loss  
630 function. For the  $k$ -th parameter, the contribution to the change in loss during task  $\mu$  is

$$\omega_k^\mu = \sum_{t=t^{\mu-1}}^{t^\mu} g_k(\theta(t)) \Delta\theta_k(t), \quad (30)$$

631 where  $g_k(\theta(t))$  is the gradient of loss with respect to  $\theta_k$  evaluated at  $\theta_k(t)$ , i.e.,  $\frac{\partial \mathcal{L}}{\partial \theta_k} |_{\theta_k(t)}$ , and  $\Delta\theta_k(t)$   
632 is the parameter change taken at step  $t$ . Therefore  $\omega_k^\mu$  tracks how parameter  $\theta_k$  contributes to



633 changes in the loss during the  $\mu$ -th task (from  $t^{\mu-1}$  to  $t^\mu$ ). The final synaptic importance is com-  
634 puted by first normalizing  $\omega_k^\mu$  with the total change in the synaptic weight  $\Delta_k^\mu = \theta_k(t^\mu) - \theta_k(t^{\mu-1})$ ,  
635 and summing  $\omega_k^v$  for all tasks  $v < \mu$ .

$$\Omega_k^\mu = \sum_{v < \mu} \frac{\omega_k^v}{(\Delta_k^v)^2 + \xi}. \quad (31)$$

636 The additional hyperparameter  $\xi$  prevents  $\Omega_k^\mu$  from becoming too large. The hyperparameters  
637  $c = 0.1$  and  $\xi = 0.01$  are determined by a coarse grid search. The final loss is the sum of the  
638 squared-error loss and the continual learning regularizer.

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{cont}}. \quad (32)$$

639 Even with the help of the continual learning technique, we had difficulties training the net-  
640 work using our original task setups. So we made the decision-making tasks easier by increasing  
641 the coherences by 4 times. We also made the delayed decision-making tasks easier by increasing  
642 the coherence by 2 times. In addition, we used the rectified linear function as the neuronal non-  
643 linearity, namely  $f(x) = \max(x, 0)$ . We found that networks using rectified linear units learned  
644 context-dependent tasks (Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2) more easily.

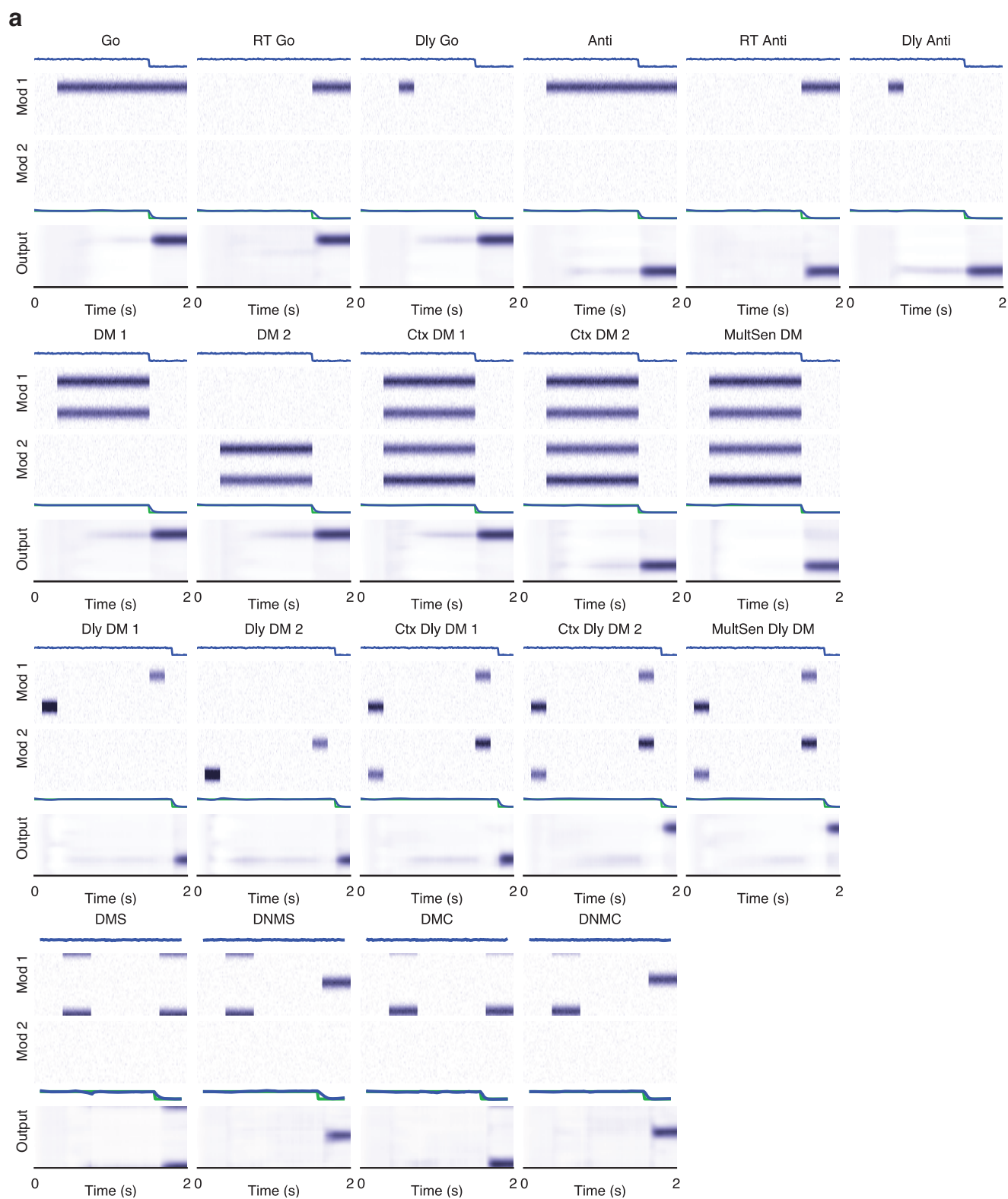
- 645 1. Fuster, J. *The Prefrontal Cortex* Fifth Edition (Academic Press, 2015).
- 646 2. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review*  
647 *of neuroscience* **24**, 167–202 (2001).
- 648 3. Wang, X.-J. in *Principles of Frontal Lobe Function* (eds Stuss, D. T. & Knight, R. T.) Second  
649 Edition, 226–248 (New York: Cambridge University Press, 2013).
- 650 4. Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode ab-  
651 stract rules. *Nature* **411**, 953–6 (2001).
- 652 5. Sakai, K. Task set and prefrontal cortex. *Annu Rev Neurosci* **31**, 219–45 (2008).
- 653 6. Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W. & Braver, T. S. Rapid transfer of abstract  
654 rules to novel contexts in human lateral prefrontal cortex. *Front Hum Neurosci* **5**, 142 (2011).
- 655 7. Tschentscher, N., Mitchell, D. & Duncan, J. Fluid intelligence predicts novel rule implemen-  
656 tation in a distributed frontoparietal control network. *J Neurosci* **37**, 4841–4847 (2017).
- 657 8. Hanes, D. P., Patterson, W. F. 2nd & Schall, J. D. Role of frontal eye fields in countermanding  
658 saccades: visual, movement, and fixation activity. *J Neurophysiol* **79**, 817–34 (1998).
- 659 9. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic  
660 value. *Nature* **441**, 223–6 (2006).
- 661 10. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature*  
662 **497**, 585–90 (2013).
- 663 11. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by  
664 recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
- 665 12. Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: a new window into the  
666 human brain's unique capacity for flexible cognitive control. *Cogn Affect Behav Neurosci* **13**,  
667 1–22 (2013).
- 668 13. Reverberi, C., G6rgen, K. & Haynes, J.-D. Compositionality of rule representations in human  
669 prefrontal cortex. *Cereb Cortex* **22**, 1237–46 (2012).
- 670 14. Zipser, D & Andersen, R. A. A back-propagation programmed network that simulates re-  
671 sponse properties of a subset of posterior parietal neurons. *Nature* **331**, 679–84 (1988).

- 672 15. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory-inhibitory recurrent neural net-  
673 works for cognitive tasks: a simple and flexible framework. *PLoS Comput Biol* **12**, e1004792  
674 (2016).
- 675 16. Carnevale, F., de Lafuente, V., Romo, R., Barak, O. & Parga, N. Dynamic control of response  
676 criterion in premotor cortex during perceptual detection under temporal uncertainty. *Neu-*  
677 *ron* **86**, 1067–77 (2015).
- 678 17. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent network models of sequence generation  
679 and memory. *Neuron* **90**, 128–42 (2016).
- 680 18. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J. & Wang, X.-J. Computing by  
681 robust transience: how the fronto-parietal network performs sequential, category-based  
682 decisions. *Neuron* **93**, 1504–1517 (2017).
- 683 19. Eliasmith, C. *et al.* A large-scale model of the functioning brain. *Science* **338**, 1202–5 (2012).
- 684 20. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the  
685 monkey's dorsolateral prefrontal cortex. *J Neurophysiol* **61**, 331–49 (1989).
- 686 21. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu Rev Neurosci* **30**, 535–  
687 74 (2007).
- 688 22. Siegel, M., Buschman, T. J. & Miller, E. K. Cortical information flow during flexible sensori-  
689 motor decisions. *Science* **348**, 1352–5 (2015).
- 690 23. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports  
691 evolving demands during decision-making. *Nat Neurosci* **17**, 1784–92 (2014).
- 692 24. Romo, R., Brody, C. D., Hernández, A & Lemus, L. Neuronal correlates of parametric working  
693 memory in the prefrontal cortex. *Nature* **399**, 470–3 (1999).
- 694 25. Munoz, D. P. & Everling, S. Look away: the anti-saccade task and the voluntary control of  
695 eye movement. *Nat Rev Neurosci* **5**, 218–28 (2004).
- 696 26. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory  
697 in prefrontal cortex of the macaque. *J Neurosci* **16**, 5154–67 (1996).
- 698 27. Freedman, D. J. & Assad, J. A. Neuronal mechanisms of visual categorization: an abstract  
699 view on decision making. *Annu Rev Neurosci* **39**, 129–47 (2016).

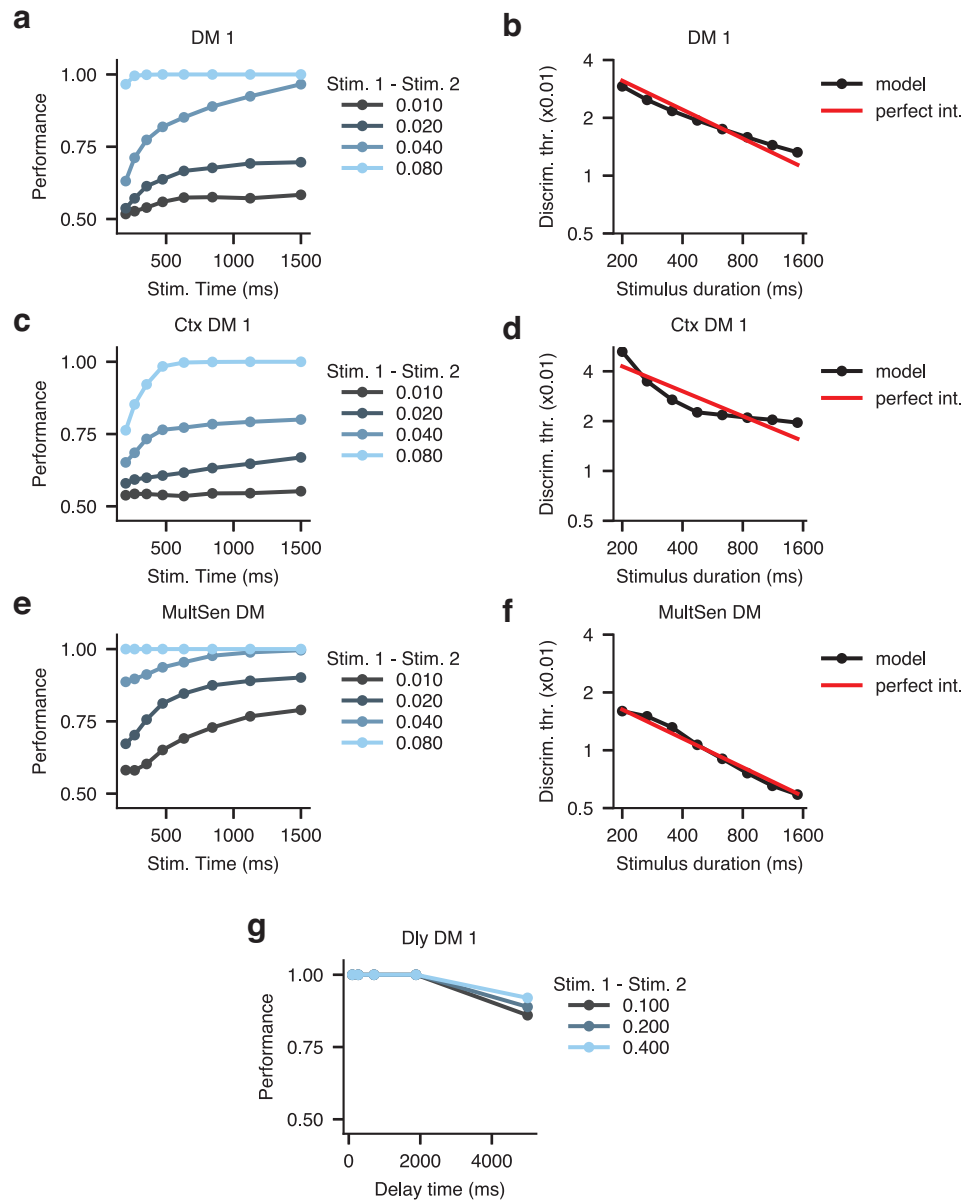
- 700 28. Abbott, L. F. & Chance, F. S. Drivers and modulators from push-pull and balanced synaptic  
701 input. *Prog Brain Res* **149**, 147–55 (2005).
- 702 29. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional  
703 recurrent neural networks. *Neural Comput* **25**, 626–49 (2013).
- 704 30. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*  
705 **36**, 955–68 (2002).
- 706 31. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of  
707 words and phrases and their compositionality. *Advances in neural information processing*  
708 *systems*, 3111–3119 (2013).
- 709 32. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad*  
710 *Sci U S A* **114**, 3521–3526 (2017).
- 711 33. Zenke, F., Poole, B. & Ganguli, S. Improved multitask learning through synaptic intelligence.  
712 *arXiv preprint arXiv:1703.04200* (2017).
- 713 34. Benna, M. K. & Fusi, S. Computational principles of synaptic memory consolidation. *Nat*  
714 *Neurosci* **19**, 1697–1706 (2016).
- 715 35. Mirabella, G. *et al.* Neurons in area V4 of the macaque translate attended visual features  
716 into behaviorally relevant categories. *Neuron* **54**, 303–18 (2007).
- 717 36. Wimmer, R. D. *et al.* Thalamic control of sensory selection in divided attention. *Nature* **526**,  
718 705–9 (2015).
- 719 37. Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J. & Fusi, S. Internal representation of task rules  
720 by recurrent dynamics: the importance of the diversity of neural responses. *Front Comput*  
721 *Neurosci* **4**, 24 (2010).
- 722 38. Cole, M. W. *et al.* Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat*  
723 *Neurosci* **16**, 1348–55 (2013).
- 724 39. Machens, C. K., Romo, R. & Brody, C. D. Functional, but not anatomical, separation of  
725 "what" and "when" in prefrontal cortex. *J Neurosci* **30**, 350–60 (2010).
- 726 40. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human  
727 prefrontal cortex. *Science* **302**, 1181–5 (2003).

- 728 41. Badre, D. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal  
729 lobes. *Trends Cogn Sci* **12**, 193–200 (2008).
- 730 42. Caruana, R. Multitask learning. *Machine Learning* **28**, 41–75 (1997).
- 731 43. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–  
732 33 (2015).
- 733 44. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses  
734 in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619–24 (2014).
- 735 45. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The distributed  
736 nature of working memory. *Trends Cogn Sci* **21**, 111–124 (2017).
- 737 46. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks  
738 for cognitive and value-based tasks. *Elife* **6**, e21492 (2017).
- 739 47. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. Soft-  
740 ware available from [tensorflow.org](http://tensorflow.org). <<http://tensorflow.org/>> (2015).
- 741 48. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*  
742 (2014).
- 743 49. Le, Q. V., Jaitly, N. & Hinton, G. E. A simple way to initialize recurrent networks of rectified  
744 linear units. *arXiv preprint arXiv:1504.00941* (2015).

745 **SUPPLEMENTARY FIGURES**

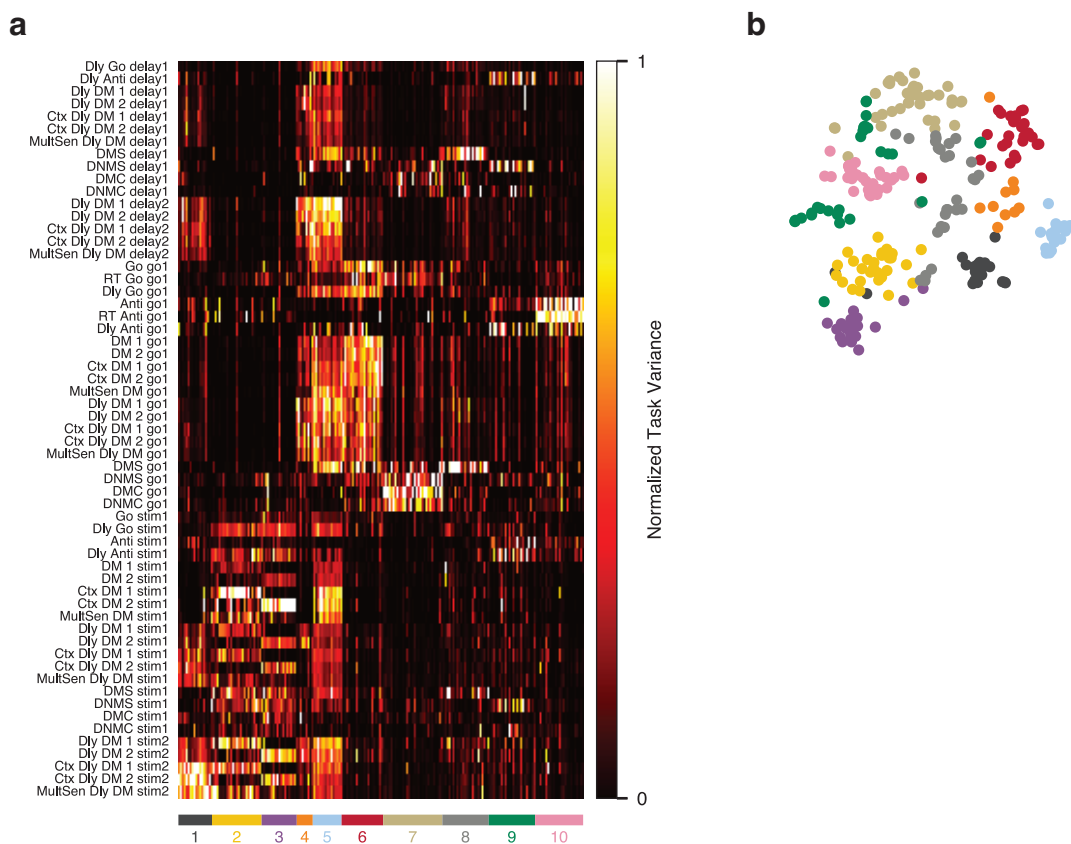


**Supplementary Fig. 1.** Sample trials from the 20 tasks trained. (a) Convention is the same as **Fig. 1a**. Output activities are obtained from a sample network after training. Green lines are the target activities for the fixation output unit.

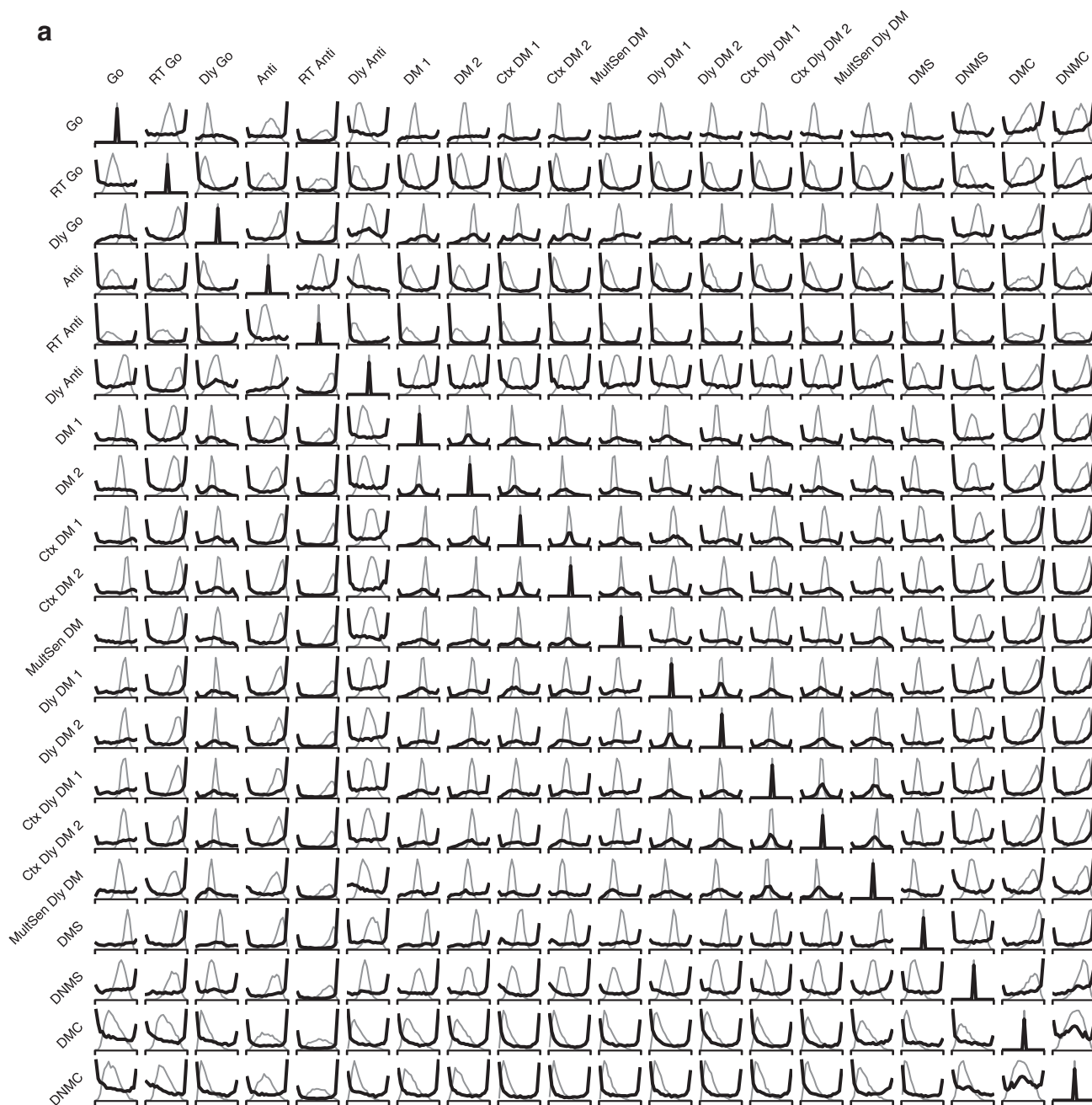


**Supplementary Fig. 2.** Psychometric tests for a range of tasks. **(a)** Decision making performances improve with longer stimulus presentation time and stronger stimulus coherence in the DM 1 task in a sample network. **(b)** Discrimination thresholds decrease with longer stimulus presentation time in the DM 1 task. The discrimination thresholds are estimated by fitting cumulative Weibull functions. **(c-f)** Same analyses as **(a,b)** for the Ctx DM 1 **(c,d)** and MultSen DM **(e,f)** task. These results are obtained from one sample network. The capability to integrate information over time varies across networks. However, this variation has no impact on other results. **(g)** A sample network is able to perform well above chance in the Dly DM 1 task for a delay period of up to five seconds.

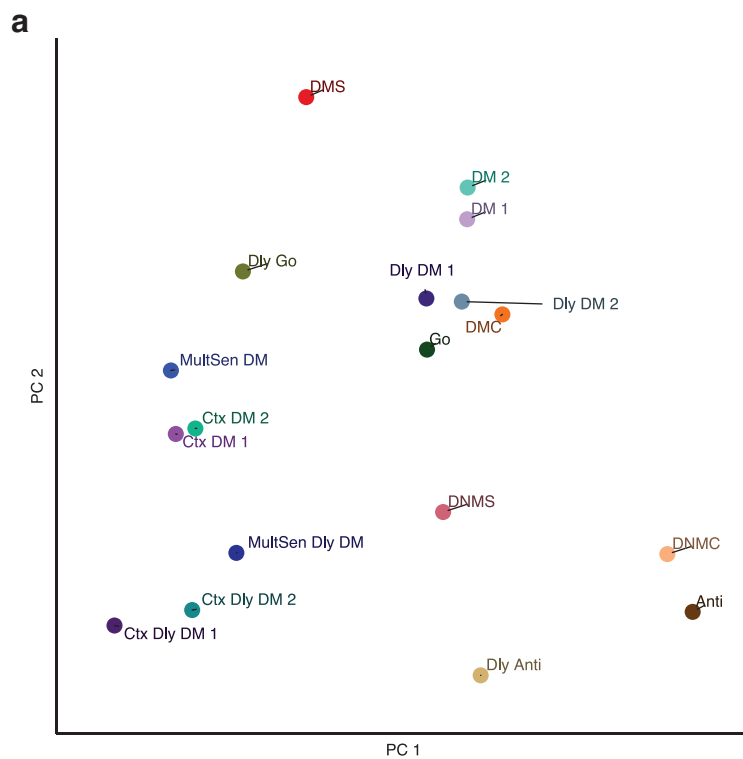




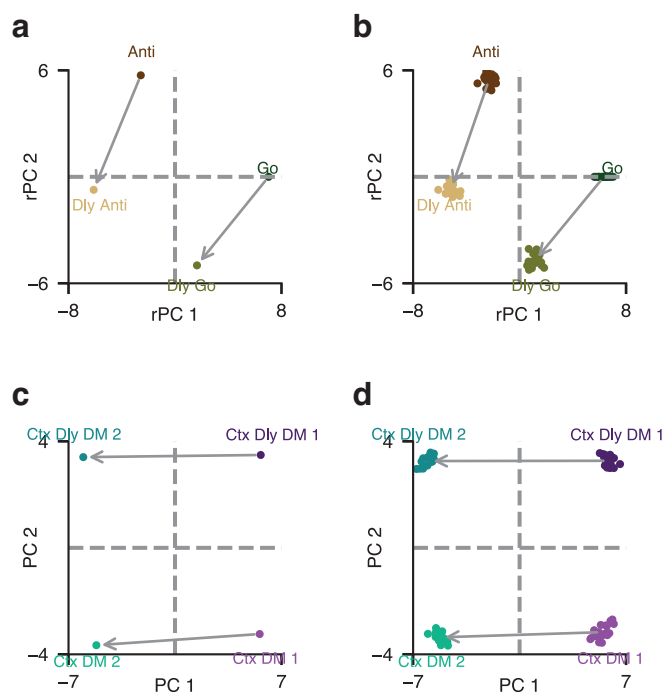
**Supplementary Fig. 3.** Epoch variances across all task epochs and active units. **(a)** Epoch variance is computed in a similar way to task variance, except that it is computed for individual task epochs instead of tasks. There are clusters of units that are selective in specific epochs. **(b)** Visualization of the epoch variance map in the same style as **Fig. 3d**.



**Supplementary Fig. 4.** Fractional variance distributions for all pairs of tasks. (a) There is a total of 190 unique pairs of tasks from all 20 tasks trained. Each fractional variance distribution (black) shown here is averaged across 20 networks. As a control, we also computed fractional variance distributions (gray) from activities of surrogate units that are generated by randomly mixing activities of the original network units (see Online Methods).



**Supplementary Fig. 5.** Representation of all tasks in state space. (a) The representation of each task is computed the same way as in **Fig. 5**. Here showing the representation of all tasks in the top two principal components. RT Go and RT Anti tasks are not shown here because there is no well-defined stimulus epoch in these tasks.



**Supplementary Fig. 6.** Visualization of connection weights of rule inputs. (a) Connection weights from rule input units representing Go, Dly Go, Anti, Dly Anti tasks visualized in the space spanned by the top two principal components (PCs) for a sample network. Similar to 5, the top two PCs are rotated and reflected (rPCs) to form the two axes. (b) The same analysis as in (a) is performed for 20 networks, and the results are overlaid. (c) Connection weights from rule input units representing Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2 tasks visualized in the top two PCs for a sample network. (d) The same analysis as in (c) for 20 networks.