# Residual dynamics resolves recurrent contributions to neural computation

Aniruddh R. Galgali[1, 2, †], Maneesh Sahani[3] and Valerio Mante[1, 2]

[1] Institute of Neuroinformatics, University of Zurich & ETH Zurich, Zurich, Switzerland
[2] Neuroscience Center Zurich, University of Zurich & ETH Zurich, Zurich, Switzerland
[3] Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom
[†]Current Affiliation : Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
Corresponding Authors: *aniruddh.galgali@psy.ox.ac.uk*, *valerio@ini.uzh.ch*

## Abstract

Relating neural activity to behavior requires an understanding of how neural computations arise from the coordinated dynamics of distributed, recurrently connected neural populations. However, inferring the nature of recurrent dynamics from partial recordings of a neural circuit presents considerable challenges. Here we show that some of these challenges can be overcome by a fine-grained analysis of the dynamics of neural residuals-that is, trial-by-trial variability around the mean neural population trajectory for a given task condition. Residual dynamics in macaque pre-frontal cortex (PFC) in a saccade-based perceptual decision-making task reveals recurrent dynamics that is time-dependent, but consistently stable, and suggests that pronounced rotational structure in PFC trajectories during saccades is driven by inputs from upstream areas. The properties of residual dynamics restrict the possible contributions of PFC to decision-making and saccade generation, and suggest a path toward fully characterizing distributed neural computations with large-scale neural recordings and targeted causal perturbations.

## Introduction

Perception, decisions, and the resulting actions reflect neural computations implemented by large, interacting neuronal populations acting in concert[1,2]. Inferring the nature of these interactions from recordings of neural activity is a key step toward uncovering the neural computations underlying behavior[3–9]. One promising approach assumes that neural computations are instantiated by a dynamical system[10,11], reflecting the combined effects of feed-forward inputs into a neural population and dynamics implemented through its recurrent connectivity[11–16]. The utility of this "computation-through-dynamics" framework hinges on our ability to disentangle how inputs and recurrent dynamics contribute to the activity of a neural population[7,17,18].

Here, we show that the properties of inputs and recurrent dynamics can sometimes be revealed by analyzing the dynamical structure of neural population residuals- that is, the trial-to-trial variability in neural population responses[19–25]. Our approach is based on the intuitive idea that the effect of recurrent computations can be revealed by observing how a perturbation of the state of the neural population evolves over time[26–29]. Unlike experiments employing external, causal perturbations, we directly analyze response residuals, which we interpret as naturally occurring perturbations within the repertoire of activity patterns produced by a recurrent neural network[30,31]. We refer to the dynamics of response residuals as "residual dynamics", and show that it provides insights into the combined effects of the recurrent dynamics implemented locally in the recorded area and in upstream areas providing inputs to it. Obtaining a complete and quantitative description of residual dynamics is difficult, because the structured component of neural population residuals is typically dwarfed by unstructured noise that may reflect variability in single-neuron spiking[19–21]. We obtain reliable, unbiased estimates of residual dynamics with novel statistical methods based on subspace identification[32,33] and instrumental variable regression[34].

Our findings are organized in three sections. First, we illustrate the challenges in disentangling inputs and recurrent dynamics based on the simulations of simple dynamical system models (Fig. 1-2). These models implement dynamics previously proposed to explain neural population responses during sensory evidence integration[12,35] and movement generation[13,36,37]. We use the simulations to establish what insights into recurrent dynamics can be obtained from different components of the neural responses, in particular condition-averaged responses and response residuals. Second, we study neural population recordings from pre-frontal cortex (PFC) of macaque monkeys during decision-making and saccadic choices (Fig. 3-5). While condition-averaged responses in PFC are consistent with a number of previously proposed models of evidence integration and movement generation, we rule out several candidate models based on the properties of the inferred residual dynamics. Third, we study simulations of multi-area, recurrent neural network (RNN) models of decision-making[38] to illustrate how inferred residual dynamics could be used to deduce circuit-level implementations of distributed recurrent computations (Fig. 6-8).

## Results

In the framework of computation through dynamics, the temporal evolution of the state of a neural population ($\mathbf{z}_t$, $t$ indicates time) can be described through a differential equation:

$$\dot{\mathbf{z}}_t = \mathbf{F}(\mathbf{z}_t) + \mathbf{u}_t + \boldsymbol{\epsilon}_t \qquad (1)$$

The momentary change in the population state ($\dot{\mathbf{z}}_t$) on each trial reflects the combined effect of four distinct factors: the *recurrent dynamics* $\mathbf{F}(\mathbf{z}_t)$, the *inputs* $\mathbf{u}_t$, the *latent noise* $\boldsymbol{\epsilon}_t$, and the initial conditions $\mathbf{z}_0$ (state at time zero). The first three factors are assumed to combine additively, as is approximately the case in many RNN models[12–15].

Mapping these factors directly onto individual brain areas (Fig. 1a, anatomical view) is typically not possible when using neural recordings from only one or few areas within a larger network[18,39,40]. Rather, here $\mathbf{z}_t$ represents a low-dimensional dynamical state that is reflected in the collective activity of all recorded neurons[31], whereby each factor contributing to it can be distributed across many areas[41] (Fig. 1a, functional view). Nonetheless, the various factors in Equation 1 can be distinguished at a functional level, through their distinct contributions to variability in neural responses—$\mathbf{F}(\mathbf{z}_t)$ captures the functional consequences of distributed recurrent connectivity and induces variability over slow time-scales (i.e. long temporal autocorrelation); $\boldsymbol{\epsilon}_t$ captures fast variability (no autocorrelation); and $\mathbf{u}_t$ can capture fast or slow variability, depending on the complexity of processing in areas upstream of the recorded one (Fig. 1b).

We illustrate the relation between the anatomical and functional interpretations by considering two simulated scenarios differing in the complexity of the inputs. Inputs are either "simple", reflecting purely feed-forward computations (Fig. 1b, top; Fig. 1c-d, 2) or "complex", resulting from recurrent processing occurring upstream of the recorded area (Fig. 1b, bottom; Fig. 6,7). These simulations illustrate the challenges in distinguishing the functional contributions of recurrent dynamics and inputs, but also that response residuals are well-suited for this challenge.

### Neural trajectories poorly constrain recurrent computations

We simulated responses of several hand-designed models that approximate neural population dynamics previously proposed to underly the accumulation of sensory evidence toward a choice[12,35] (Fig. 1c) or the generation of complex motor sequences[13,37] (Fig. 1d). As in more complex RNN models[12,3513,37], here the input consists of two components (Fig. 1b, functional view): a deterministic input drive $\bar{\mathbf{u}}_t$ (repeatable across trials of the same condition) and latent input noise $\tilde{\boldsymbol{\epsilon}}_t$ (Fig. 1b, simple inputs).

We simulated single-trial responses for two task-conditions and visualized them as trajectories

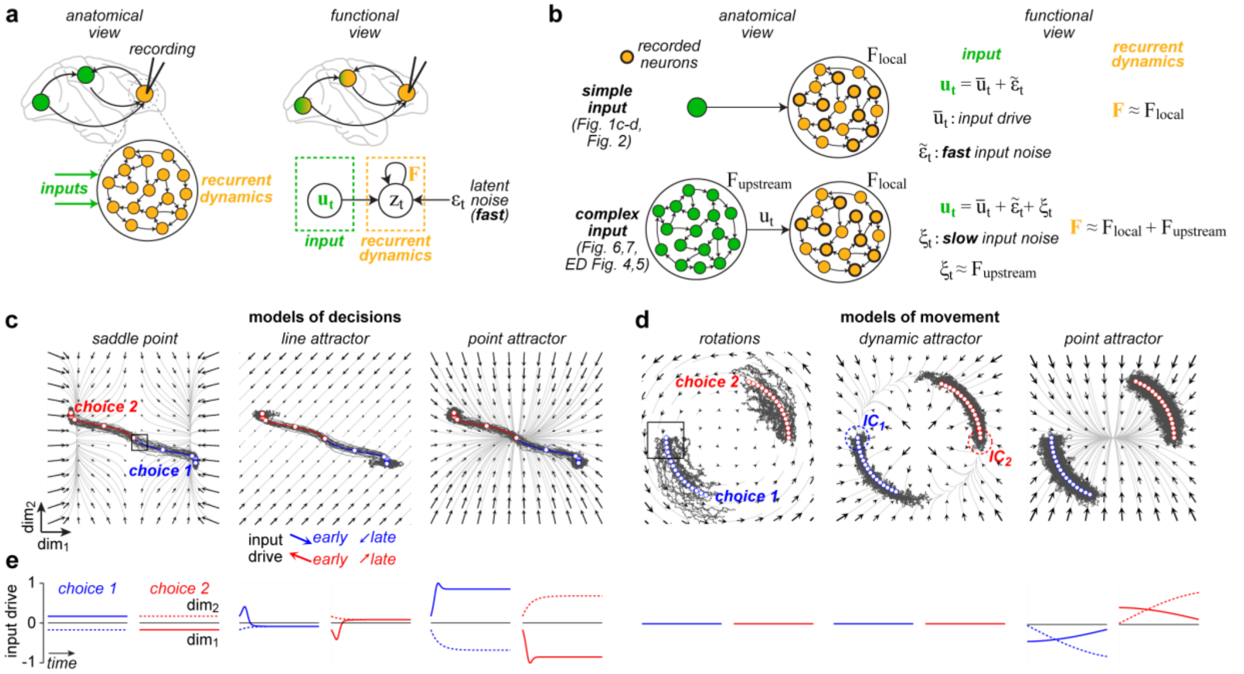


***Figure 1. Disentangling contributions of inputs and recurrent dynamics to neural responses.***

***a**, Computation through dynamics. Anatomical view (left): recurrent dynamics and inputs respectively capture how the recorded neural responses are shaped by recurrent connectivity within the recorded area (orange) and by responses in additional areas (green). Functional view (right): recurrent dynamics and inputs reflect processes distributed across several areas (color gradient) and are defined based on their functional contributions to neural responses (graphical model, bottom). **b**, Relation of functional and anatomical viewpoints in two example scenarios (top & bottom row: simple vs. complex inputs). **c-d**, Models of decision-making (**c**) and movement generation (**d**) based on simple inputs as in b (top). Each panel shows simulated single trials (dark-gray trajectories) and condition-averaged trajectories (blue and red trajectories) for two task conditions (choice 1 and 2). Black arrows show the effect of recurrent dynamics on the response at any state-space location. The effect of an input drive is constant across state-space, but can change over time and across task conditions (middle panel in **c**, example input directions at bottom). **c**, Models of decision-making. The three models implement unstable (left), perfect (middle), and leaky integration (right) of an appropriately chosen input. **d**, Models of movement-generation. Left: purely rotational dynamics. Perturbations along both state-space dimensions are persistent; Middle: dynamic attractor. Perturbations along the radial dimension decay, but persist along the circular "channel". Right: point attractor. Responses are driven by strong inputs. IC: approximate extent of the initial conditions, shown for the dynamic attractor model. **e**, Input drive (see **b**) for the models in **c** and **d**. Curves indicate the components of the input drive along the two state-space dimensions (solid vs dashed) over time (horizontal axis) and conditions (red vs blue). Input drives are chosen to produce identical condition-averaged trajectories across models in **c**, and in **d**. Boxes in **c** and **d** (left sub-panels): regions of state-space analyzed in Fig. 2.*

in a two-dimensional (2D) neural state-space (Fig. 1c,d, choice 1 & 2; dark-gray curves). The recurrent dynamics $\mathbf{F}(\mathbf{z}_t)$ describes the noiseless evolution of the instantaneous state ($\mathbf{z}_t$) from a given state-space location in the absence of inputs (Fig. 1c,d, black arrows and light-gray curves). The input drive ($\bar{\mathbf{u}}_t$) injects a particular pattern of activity into the neural population, thus pushing the state along a state-space direction that could vary across time and task conditions (Fig. 1c; red and blue arrows, and Fig. 1e).

Very different combinations of recurrent dynamics and inputs resulted in very similar trajectories. The three models of decision-making instantiate different behavioral ‘strategies’ for perceptual decision-making[42], from unstable, impulsive decisions (Fig. 1c, saddle point), to optimal accumulation of evidence (Fig. 1c, line attractor), and leaky, forgetful accumulation (Fig. 1c, point attractor). Yet, for the chosen input drive, which can be constant (Fig. 1e saddle point) or transient (Fig. 1e, line and point attractor), all three models produce similar single-trial trajectories (Fig. 1c, gray curves) and indistinguishable condition-averaged trajectories (Fig. 1c, blue and red curves). Analogous observations hold for the models of movement generation (Fig. 1d). The condition-averages do not distinguish between two models in which responses were driven solely by recurrent dynamics (Fig. 1e) – a model implementing rotational dynamics[13,36], in which variability in the initial condition is reflected throughout the entire trajectory (Fig. 1d rotations; gray curves), and a “dynamic attractor”[37] model, in which activity is pushed towards and through a narrow channel in state space (Fig. 1d, dynamic attractor). The resulting condition-averages are also identical to those from a model that implements point-attractor recurrent dynamics and is strongly input driven[18] (Fig. 1d, point attractor).

Condition-averaged trajectories, which are often used to compare simulated neural responses to measured population activity[12,13,43], thus cannot disentangle the functional effects of recurrent dynamics and inputs in these simple models.

### Residual dynamics can resolve recurrent contributions

Neural residuals are defined as the difference between a single-trial trajectory and the corresponding condition-averaged trajectory[20,44] (Extended Data Fig. 1). We interpret residuals as perturbations away from the condition-averaged trajectory, and capture how these perturbations evolve over time through the “residual dynamics” (Extended Data Fig. 1).

For the simulated models, the dynamics of residuals can be derived analytically, in two steps (Fig. 2a, Extended Data Fig. 1). We define the *effective dynamics* by summing the contribution of recurrent dynamics and input drive, thus capturing the noiseless evolution of the population response from any given state-space location. We then obtain the *residual dynamics* by subtracting, from the effective dynamics, a component corresponding to the instantaneous direction of change along the condition-averaged trajectory (Fig. 2a, see labels over each panel).

The residual dynamics describes how a perturbation of a neural state along the condition-averaged trajectory evolves over the course of one time-step (Fig. 2c,d, blue dot: unperturbed “reference” neural state; arrows: evolution from the perturbed states). For the saddle point model (Fig. 2c, saddle point), perturbations along the horizontal direction expand over time (arrows point away from the reference state), whereas perturbations along the vertical direction

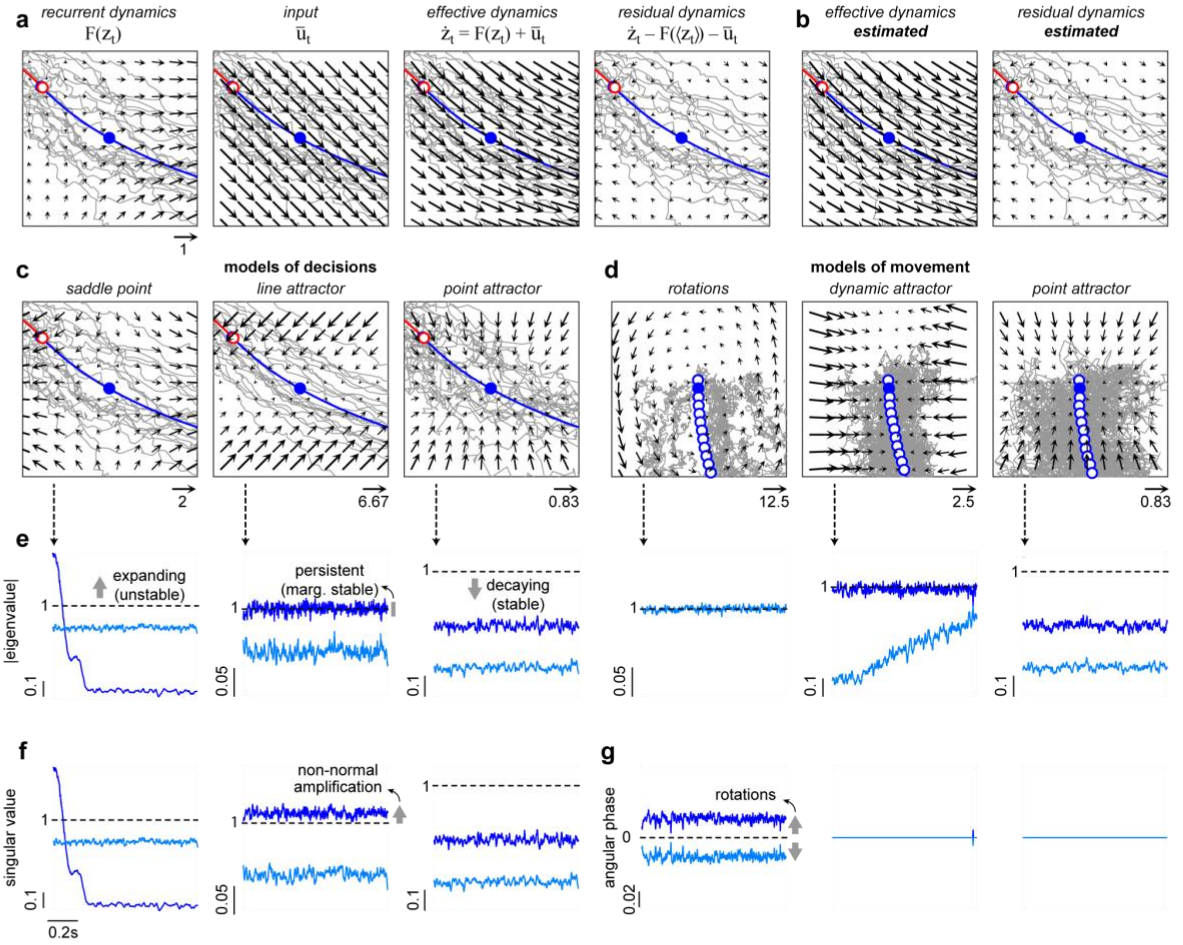


*Figure 2. Residual dynamics reveals population-level computations.*

***a**, Different factors contributing to the dynamics of the saddle point model, shown in the state-space region marked in Fig. 1c for an early time in choice 1 trials (box). Same conventions as in Fig. 1c. Recurrent dynamics and input drive sum to generate the effective dynamics, determining the evolution of the response in the absence of noise. The residual dynamics is the component of the effective dynamics that explains the evolution of perturbations away from the condition-averaged trajectory (blue line; blue dot: reference time). **b**, Effective and residual dynamics estimated directly from simulated single-trial residuals match the ground-truth in **a**. **c**, Ground-truth residual dynamics for the models of decisions, same state-space region and reference time as in **a**. The residual dynamics reflects the key properties of the recurrent dynamics at the corresponding state-space region in Fig. 1c. The arrows in each flow field were scaled by a fixed factor that differed across models and with **a** (numbers close to arrows at the bottom). **d**, Analogous to **c**, but for the models of movement at an early time in choice 1 trials (box in Fig. 1d). **e**-**g**, Properties of the estimated residual dynamics for the models in Fig. 1c-d. Only residual dynamics for choice 1 is shown. The residual dynamics is described by a time and condition-dependent, autonomous, linear dynamical system. The corresponding time-varying dynamics matrices describe the residual dynamics at particular locations along one of the condition-averaged trajectories (Extended Data Fig. 1). **e**, Magnitude of the eigenvalues (EV, y-axis) of the 2D dynamics matrix as a function of time (x-axis). **f**, Singular values (SV) of the dynamics matrix as a function of time for the models of decisions. The difference between EV and SV in the line-attractor model is a consequence of non-normal dynamics. **g**, Angular phase associated with complex-valued EV for models of movement. Larger angular phase implies faster rotational dynamics. EVs, SVs, and angular phase together distinguish between the different models.*

decay back to the trajectory (arrows point towards the reference state). These dynamics correctly reflect the influence of a saddle point in the vicinity of the reference state (Fig. 1c, box). Likewise, the residual dynamics correctly reveals line attractor and point attractor dynamics in the other two models of decisions (Fig. 2c) and key properties of the recurrent dynamics in the models of movement, i.e. rotational dynamics, decay towards the dynamic attractor, and point attractor dynamics (Fig. 2d). These differences in the underlying recurrent dynamics are less apparent in the effective dynamics, particularly for strong input drives (Extended Data Fig. 1).

For measured neural responses, we approximate residual dynamics with a condition and time-dependent, locally linear system, whereby time parameterizes location in state-space along the condition-averaged trajectory (Extended Data Fig. 1). Such linear dynamics is well-suited to describe residuals because, by definition, residual dynamics always has a fixed point at the location of the reference neural state (Fig. 2c,d, blue dot). We estimate the linear approximations by combining methods from subspace identification[33,45] and instrumental variable regression[34] (Extended Data Fig. 2). These methods, unlike simpler linear regression approaches, can produce robust and unbiased estimates of residual dynamics in biologically realistic settings (Extended Data Fig. 3).

We summarize the residual dynamics through three properties of the linear approximations, specifically the magnitude of the eigenvalues (EV), the singular values (SV), and the rotation frequency associated with the EV (Fig. 2e-g). Together, these properties distinguish the models in Fig. 1c-d. For locations close to the saddle point in the model of decision-making, one EV is larger than 1, implying that perturbations along the associated eigenvector (the horizontal direction in Fig. 1c, left) *expand* over time; the other EV is smaller than one, corresponding to *decay* along the vertical direction (Fig. 1c, left; center of flow field; Fig. 2e, left-most panel; early times). For the line attractor, the largest EV is 1 (Fig. 2e, second from left) as horizontal perturbations are *persistent* - that is, neither expand nor decay. For a point attractor, all EV smaller than 1 (Fig. 2e, third from left; all directions decay). Rotational dynamics results in complex-valued EV associated with a non-zero rotation frequency (Fig. 2g). Differences between the magnitude of SV and EV reflect non-normal dynamics, a feature of many models of neural computation[46–48]. The SV larger than 1 in the line attractor model implies that small perturbations along the corresponding right singular vector transiently expand, even though they are persistent (EV=1) or decay (EV<1) over longer time-scales (Fig. 2e,f).

### Residuals dynamics reflects local and upstream recurrence

The above simulations illustrate one setting in which residual dynamics, unlike the condition-averaged trajectories, can reveal the properties of the recurrent dynamics—when input variability is temporally uncorrelated, any slow correlations in the residuals are entirely due to (and can be used to infer) the recurrent dynamics (Fig. 1b, top; simple inputs). This constraint, however, is likely violated for single areas in biological networks, where the input into an area could result from recurrent processing in upstream areas[38,41]. In Equation 1, the input ($\mathbf{u}_t$) would then include a component of variability with slow temporal correlations, reflecting the upstream recurrent dynamics ($\boldsymbol{\xi}_t$ in Fig. 1b, bottom; complex input).

In such settings, residual dynamics reflects not just the "local" recurrent dynamics ($\mathbf{F}_{\text{local}}$, Fig. 1b), but rather the combined effects of the recurrent dynamics in the recorded area and in any upstream areas contributing an input to the recorded area[44] ($\mathbf{F}_{\text{upstream}}$, Fig. 1b). For example, residual dynamics with large EV or large rotation frequencies need not imply that the recurrent dynamics in the recorded area is unstable or rotational, as such dynamics may be implemented also, or exclusively, in areas upstream of the recorded one (Extended Data Figs. 4-5).

Notably, direct or indirect connections from unrecorded to recorded neurons within the local, recurrently connected population need not result in a functional "input" in the sense of Equation 1. If neural activity evolves within a low-dimensional manifold, recordings from a large enough subset of neurons within a network can be sufficient to estimate the population state $\mathbf{z}_t$ of the entire network[30,31]. The effect of unrecorded neurons in the local network is then fully captured by the recurrent dynamics $\mathbf{F}$ [49] (Fig. 1b, $\mathbf{F} \approx \mathbf{F}_{\text{local}}$).

## Neural trajectories of decisions and movements in PFC

We developed an analysis pipeline to estimate residual dynamics from recorded neural responses (Extended Data Fig. 2) and applied it to recordings from pre-frontal cortex (PFC; area 8Ar) in two macaque monkeys performing a saccade-based perceptual decision-making task[50] (Fig. 3a-b). We increased the statistical power of our analyses by "aligning" and combining neural activity from different experiments with a similar task-configuration (Extended Data Fig. 2, Step 1; 14-61 experiments per configuration; 150-200 units per experiment). The alignment yielded a 20-dimensional (20D) activity subspace explaining >90% of task-related variance in the average neural responses[31] (Extended Data Fig. 6). We performed subsequent analyses within this aligned subspace, although the main results can be reproduced from sufficiently long single experiments (Extended Data Fig. 7).

We visualized the aligned population trajectories through projections onto several two-dimensional activity subspaces: a "choice" plane, emphasizing choice-related activity; a "time" plane, emphasizing time-varying activity common to all conditions; and two "jPC" planes[36], emphasizing rotational dynamics (Fig. 3c,d; left to right). Only the two jPC planes were orthogonalized with respect to each other, meaning that some planes captured shared components of the activity (e.g. Fig. 3c, time and $\text{jPC}_{12}$ planes). We estimated the planes

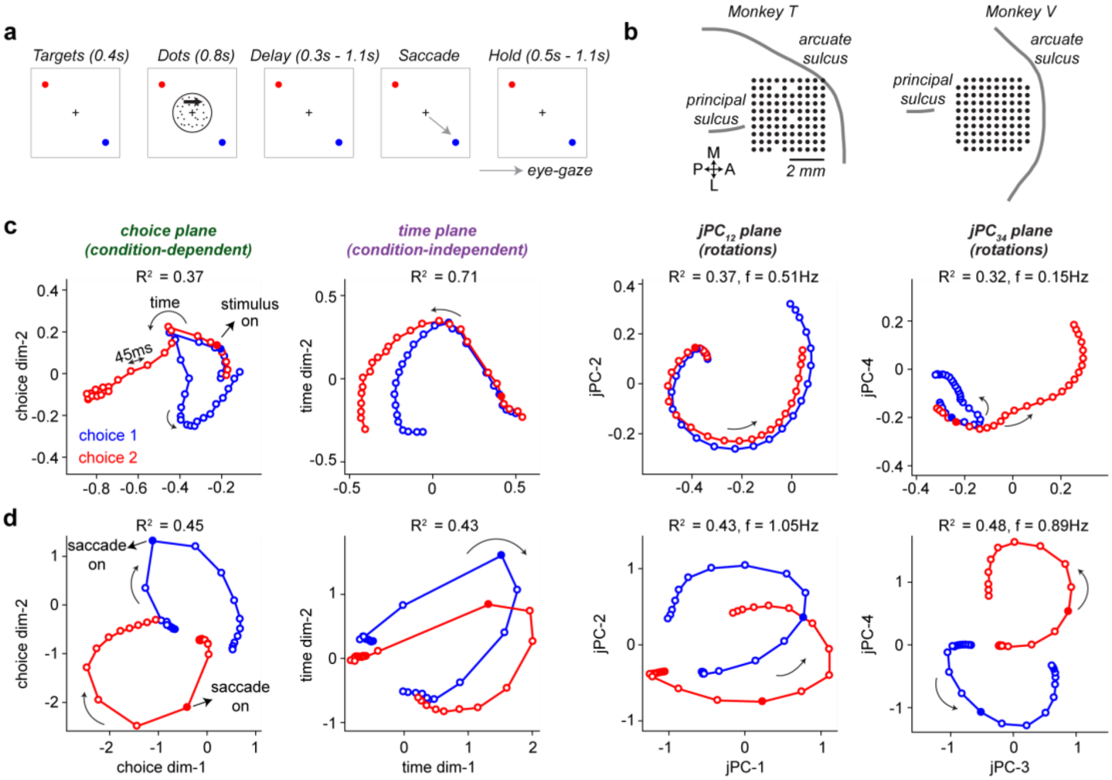


***Figure 3. Average dynamics in prefrontal cortex during perceptual decisions and saccades.***

***a**, Behavioral task. Monkeys fixating at the center of a screen (fixation point, black cross) viewed a random dot stimulus for 800ms. After a delay period of random duration, they reported the perceived direction of motion with a saccade to one of two targets (red and blue circles; blue: choice 1; red: choice 2). Following the saccade, the monkeys had to fixate on the chosen target during a hold period of random duration. **b**, Position of the 10 x 10 electrode array in pre-arcuate cortex of the two monkeys. Black circles indicate the cortical locations of the 96 electrodes used for recordings. **c-d**, Neural trajectories in monkey T, averaged over trials of the same choice. Trajectories are obtained after aligning neural responses (see Extended Data Fig. 6) from experimental sessions with a similar configuration of saccade targets (config-3, Extended Data Fig. 6). Aligned responses are projected into four activity-subspaces: the choice, time, $jPC_{12}$, and $jPC_{34}$ planes, capturing variance due to choice, time, and rotations, respectively ($R^2$: fraction of variance explained; f: rotation frequency associated with the jPC plane). **c**, Trajectories in the decision-epoch (-0.2 to 1s relative to stimulus onset, filled circle). **d**, Trajectories in the movement-epoch (-0.7 to 0.5s relative to saccade onset, filled circle).*

separately during a decision epoch (Fig. 3c; random-dots presentation) and a movement-epoch (Fig. 3d; saccade-execution).

The PFC trajectories shared several features with the model trajectories in Fig. 1c-d. As in the decision models (Fig. 1c), PFC responses started in an undifferentiated state prior to stimulus onset (Fig. 3c; choice plane; filled dots mark stimulus onset) and gradually diverged based on the upcoming choice (Fig. 3c, red vs. blue). Prior to saccade-onset, PFC responses fell into largely stationary, choice-dependent states and then transitioned into rotational dynamics following the presentation of the go cue (Fig. 3d, jPC planes), similar to the movement models (Fig. 1d).

Several features of the PFC trajectories were not reproduced by the models, including strong condition-independent components[26,28,43,51] (e.g. Fig. 3c,d, time plane), choice-related activity along multiple state-space directions (Fig. 3c, choice plane), rotational dynamics within multiple subspaces (Fig. 3c,d; jPC planes) and rotational dynamics during the decision epoch (Fig. 3c, jPC planes). These shortcomings, however, are common to all models and do not provide a basis to favor one model as an explanation of PFC responses.

### Residual dynamics in PFC

To better resolve the contributions of recurrent dynamics to the recorded responses, we characterized residual dynamics in PFC. We first estimated a "dynamics subspace", contained within the previously defined aligned subspace (Fig. 4a, Extended Data Figs. 2,6-8). The dimensions of the dynamics subspace were chosen for their ability to predict "future" residual states from "past" ones, but are well aligned with dimensions explaining task-related variance (Fig. 4a, largest dot products at small values along y-axis; Extended Data Figs. 6, 7). We estimated residual dynamics within the 8-dimensional dynamics subspace with the same approach as for the simulated models (Fig. 2e-g, Extended Data Fig. 2,8,9). Dimensions orthogonal to the dynamics subspace were associated with an EV of zero—perturbations along these directions are predicted to completely decay within one time step.

EV magnitudes were strongly time-dependent (Fig. 4b, all EV), but consistently smaller than 1 (Fig. 4e, largest EV; monkey T: $p<0.005$ for all time points; monkey V: $p < 0.01$ for 43 of 44, and $p<0.005$ for 41 of 44 time points; one-sample, single-tailed t-test , $n = 8$, 2 choices x 4 configurations) implying stable, decaying dynamics. The largest EV were associated with decay time-constants in the range 187-745ms during the decision period (0s to +0.8s following stimulus onset) and 110-913ms during the delay period (-0.5s to +0.3s relative to saccade onset) for monkey T (95th percentile CIs, medians = 352ms and 293ms, n =144, 2 choices x 4 configurations x 9 times; Fig. 4e, top), and 309-1064ms and 192-3586ms for monkey V (95% CI, medians = 489ms and 491ms, n = 144; Fig. 4e, bottom). Concurrently with the saccade onset, the largest EV consistently underwent a strong contraction (Fig. 4e; $p<3\cdot10^{-5}$ and $p<3\cdot10^{-7}$ in monkeys T and V; $H_0$: largest EV equal at -275ms vs. -5ms relative to saccade onset; two-sample, single-tailed t-test, n = 8). The largest measured time constants at saccade onset fell to median values of 159ms in monkey T and 310ms in monkey V, implying that perturbations away from the average trajectory fall back to the trajectory more rapidly during movement.

The residual dynamics had rotational components in both monkeys. In monkey T, the largest rotation frequencies in the residuals (Fig. 4g top; ≈0.5-1 Hz) lay in the approximate range of frequencies for rotations in the condition-averages (Fig. 3c,d; values for f). In monkey V, even the largest rotation frequencies in the residuals (Fig. 4g bottom, ≈0.25-0.5 Hz) were smaller than those in the condition-averages (0.71-0.84Hz, decision epoch; 1.16-1.34Hz, movement epoch;

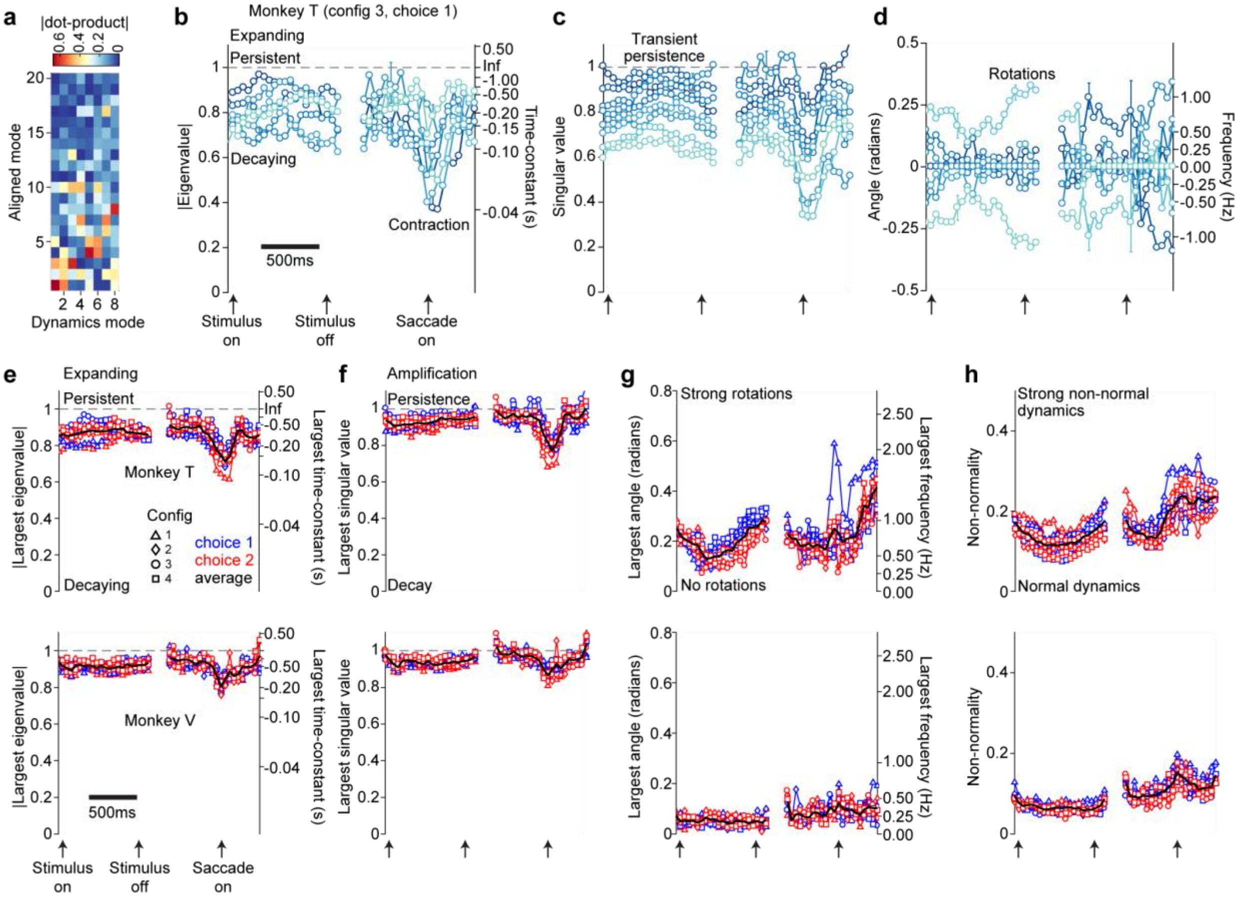


***Figure 4. Residual dynamics in prefrontal cortex during perceptual decisions and saccades.***

***a-d**, Estimated residual dynamics in prefrontal cortex in monkey T, same task configuration as in Fig. 3c,d. The residual dynamics was 8-dimensional for this example dataset. **a**, Relative alignment between the modes spanning the 8d-dynamics subspace and the modes spanning the 20d-aligned subspace (see Extended Data Figs. 6,7), measured as the absolute value of the corresponding dot-product. The dynamics modes project strongly onto the first few aligned modes, which capture most of the task-relevant variance in the responses. **b-d**, Properties of the residual dynamics (circles) for a single choice condition (choice 1). Error bars: 95% bootstrap confidence intervals (shown at selected times) obtained by fitting residual dynamics to randomly resampled trials (n = 1000). **b**, Eigenvalues (EV) of the dynamics (left axis), and associated time-constants of decay (right axis) as a function of time (x-axis). **c**, Singular values (SV) of the dynamics. The eigenvectors and singular vectors associated with the shown EV and SV can vary over time. **d**, Angular phase of the EV (left axis; angular phase = 0: real-valued EV) and associated rotation frequencies (right axis). Line colors reflect the magnitude of the EV or SV at the onset of the decision epoch. At later times, colors match those associated with the closest eigenvector or right singular vector at the preceding time. **e-h**, Properties of the residual dynamics across all animals (monkey T, top; monkey V, bottom), choices (blue: choice 1; red: choice 2), and task configurations (markers; see legend of Extended Data Fig. 6). Black curves: averages across all choices and configurations. **e**, Magnitude of the largest EV (left axis) and the associated decay time-constants (right axis). **f**, Largest singular value. **g**, Largest angular phase of the EV and the corresponding frequency of rotation. **h**, Time course of the index of non-normality.*

range across all task configurations). The largest SV of the residual dynamics exceeded the magnitude of the largest EV in both monkeys (Fig. 4e,f; $p<0.05$ for 43 of 44 and 33 of 44 timepoints in monkeys T and V; two sample, single-tailed t-test, n = 8) implying that dynamics was weakly non-normal (Fig. 4h). The largest SV were mostly smaller than 1 in both monkeys (Fig. 4f; $p<0.05$ for 41 of 44 time points in both monkeys T and V; one-sample, one-tailed t-test, $n = 8$). The non-normality is thus not sufficiently pronounced to amplify perturbations, but rather only transiently slows their decay (Fig. 4c, "transient persistence").

These findings rule out several models of recurrent dynamics. In the decision epoch, the EVs are inconsistent with unstable dynamics (EV>1, Figs. 1c,2e; saddle point) and mostly smaller than expected for persistent dynamics ($EV\approx1$, Figs. 1c,2e; line attractor). In the movement-period, the small EV around saccade onset are inconsistent with purely rotational dynamics or a dynamic attractor, which would both result in directions with slower decay ($EV\approx1$, Figs. 1d,2e; rotations and dynamic attractor). Around saccade onset (-200 to +200ms from onset), the largest EV magnitude (0.80 and 0.88 in monkeys T and V; mean, $n = 8$) and the largest rotation frequency (0.74 and 0.33 Hz in monkeys T and V; mean, $n = 8$) imply that perturbations decay by at least 50% within every 1/10th (monkey T) and 1/12th (monkey V) of a rotational cycle. During the same time window, the condition-averaged trajectories undergo about 1/4th of a rotational cycle without obvious decay. The quickly decaying residual dynamics, and the mismatch between its properties and those of the condition-averaged trajectories, are consistent with a strong input drive (Figs. 1d, 2e; point attractor).

## Alignment of residual dynamics and neural trajectories

Additional insights into how recurrent dynamics and inputs contribute to the observed activity can be gained by analyzing the inferred eigenvectors of the residual dynamics. When inputs are weak, the trajectories mostly reflect the properties of the recurrent dynamics, which in turn results in distinct relations between trajectories and eigenvectors.

We illustrate such relations in two models, obtained by augmenting the line-attractor and rotation models (Fig. 1c-d) with two new dimensions, along which recurrent dynamics was quickly decaying and input drive was strong and condition-independent. We defined activity subspaces as in Fig. 3 (Fig. 5a,c) and analyzed how they align with the eigenvectors of the residual dynamics. For the augmented line-attractor model, the choice plane is preferentially aligned (angle close to 0) with eigenvectors associated with large EV magnitudes (Fig. 5b top), as slow dynamics along these eigenvectors underlies the observed choice-related activity. For the augmented rotations model, the $jPC_{34}$ plane is preferentially aligned with the eigenvectors associated with large rotational frequencies (Fig. 5d top), as these eigenvectors underlie the rotational activity in the $jPC_{34}$ plane. Critically, the augmented subspaces are not preferentially aligned with the slow or rotational eigenvectors, as activity within them is mostly input driven. We summarize these relations with a linear regression analysis, whereby negative regression

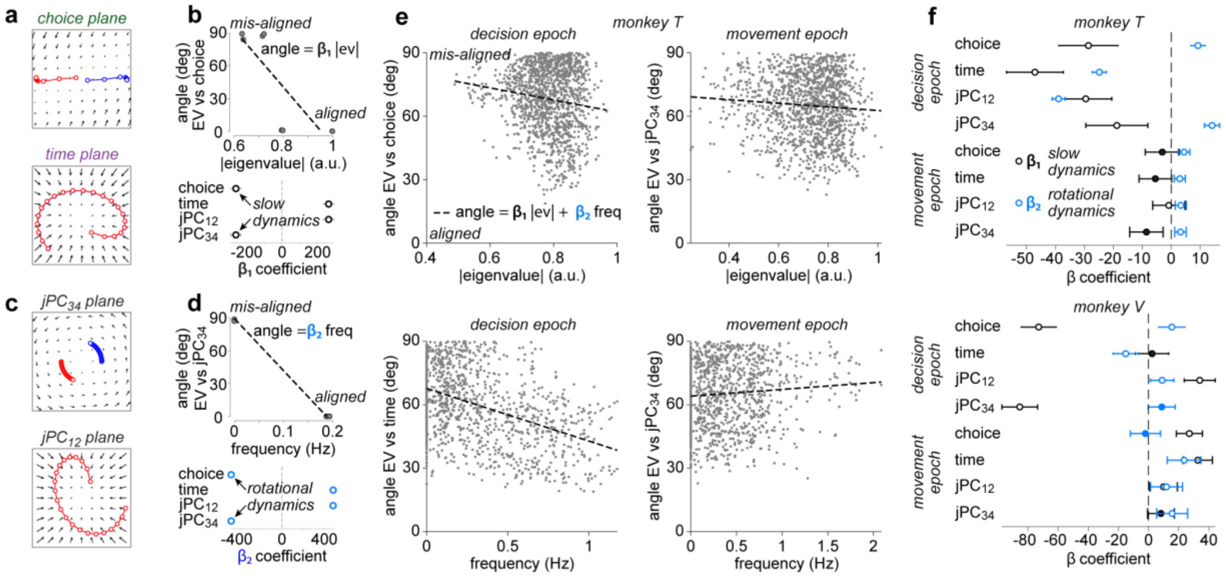


*Figure 5. Alignment of residual dynamics and condition-averaged trajectories.*

***a**, Condition-averaged trajectories for the line attractor model (top; and Fig. 1c) augmented with an additional two-dimensional subspace with decaying dynamics and strong input drive (bottom, time plane; red and blue trajectories are overlayed). **b**, Alignment between the eigenvectors of the residual dynamics and the task-related subspaces, for the model in **a**. Top: Angle between the choice plane and the eigenvectors (gray points). Eigenvectors are indexed by EV magnitude. Bottom: regression coefficients, for linear regression as on top (line; angle vs. EV magnitude). Large negative coefficients identify task-subspaces aligned with slow residual dynamics. Task-subspaces are redundant (e.g. choice and jPC$_{34}$) as residual dynamics is only 4-dimensional. **c**, Trajectories for the rotation model (top; and Fig. 1d) augmented as in **a** (bottom, jPC$_{12}$ plane). **d**, like **b**, for the model in **c**. Eigenvectors are indexed by the associated rotation frequency. Large negative coefficients identify task-subspaces aligned with rotational residual dynamics. **e**, Example alignments for PFC activity in monkey T. Angles (gray points) are pooled across times within an epoch (titles), task configurations, and choices. Linear regression (dashed line) includes coefficients ($\beta_1$ and $\beta_2$) for EV magnitude (|ev|) and rotation frequency (freq). **f**, Regression coefficients for PFC activity (as in **e**), for all epochs, task-subspaces, and monkeys (circles: coefficient estimate; error bars : 1.96 standard error). Filled circles indicate non-significant regression coefficients ($p>0.05$; t-statistic, dof = 1405, two-sided; $H_0$ : coefficient estimate = 0).*

coefficients identify planes where slow or rotational recurrent dynamics may contribute to the observed trajectories (Fig. 5b,d bottom; regression with EV magnitude or rotational frequency). The augmented, input driven subspaces in the models are, instead, aligned with fast or weakly rotational eigenvectors, resulting in positive regression coefficients (Fig. 5b,d bottom). Such positive coefficients are a trivial consequence of the low dimensionality of these models (e.g. mis-alignment with the choice plane necessarily implies alignment with the time plane) and need not occur in PFC dynamics.

We applied this analysis to PFC responses and found significant, negative coefficients primarily in the decision epoch, whereby planes containing choice-related activity were aligned with slow residual dynamics in monkeys T and V (Fig. 5f; choice and jPC$_{34}$ planes; Fig. 5e top) and rotational residual dynamics was aligned with planes containing condition-independent activity in monkey

T (Fig. 5f top, time and $jPC_{12}$ planes; Fig. 5e bottom). Coefficients in the movement epoch were mostly very small or not significant (Fig. 5f). These relations suggest that recurrent dynamics contributes to observed choice-related activity (in both monkeys) and condition-independent activity (in monkey T), but only during the decision period. Activity at the time of the saccade appears more consistent with the influence of a strong input drive[18], as we also concluded based on the quickly decaying residual dynamics in this epoch (Fig. 4e).

## Resolving local and long-range recurrence

Residual dynamics within an area can reveal key functional properties of the recurrent dynamics contributing to measured population activity, but cannot distinguish local and upstream recurrent contributions (Fig. 1b, Extended Data Fig. 4). Below, we show in simulations how such contributions could be distinguished with “global” recordings from multiple areas or by combining local recordings and causal perturbations (Fig. 7).

We simulated activity in RNNs composed of two areas (PPC and PFC, Fig. 6), characterized by local recurrence within areas and long-range connections between areas[38]. In the RNNs, PPC is upstream of PFC, as it alone receives an input with temporally uncorrelated variability (Fig. 1b, simple input) that directly encodes the external stimulus. Local recurrence is equally strong in both areas. When present, feedback connections from PFC to PPC have equal strength as the feedforward connections.

Simulated model responses in a perceptual decision-making task have choice-dependent and condition-independent components in both areas (Fig. 6a,d; choice and time modes). The EV of the residual dynamics, estimated *locally* in PPC or PFC, are typically time-dependent (Fig. 6b,e), as the RNNs are nonlinear. In particular, dynamics can change from stable ($EV<1$) to unstable ($EV>1$) after onset of the external input to PPC. We summarize the residual dynamics in each area with the peak magnitude of the EV along the corresponding choice modes (Fig. 6c,f). The choice modes define the “communication subspace” between PPC and PFC in these networks[38,44].

The simulations show that very different combinations of local and long-range connectivity can result in responses that are virtually indistinguishable based on condition-averages (Fig. 6a,d) and residual dynamics (Fig. 6b,e) computed locally. In networks with a weak feedforward connection from PPC to PFC, and no feedback from PFC, the local residual dynamics depends only (PPC) or mostly (PFC) on the strength of the local recurrence, whereby the largest EV gradually increases with stronger local recurrent connectivity (Fig. 6c). In networks with strong feedback from PFC, the local residual dynamics in both areas instead reflects the combined effects of local recurrence and long-range connectivity (Fig. 6f).

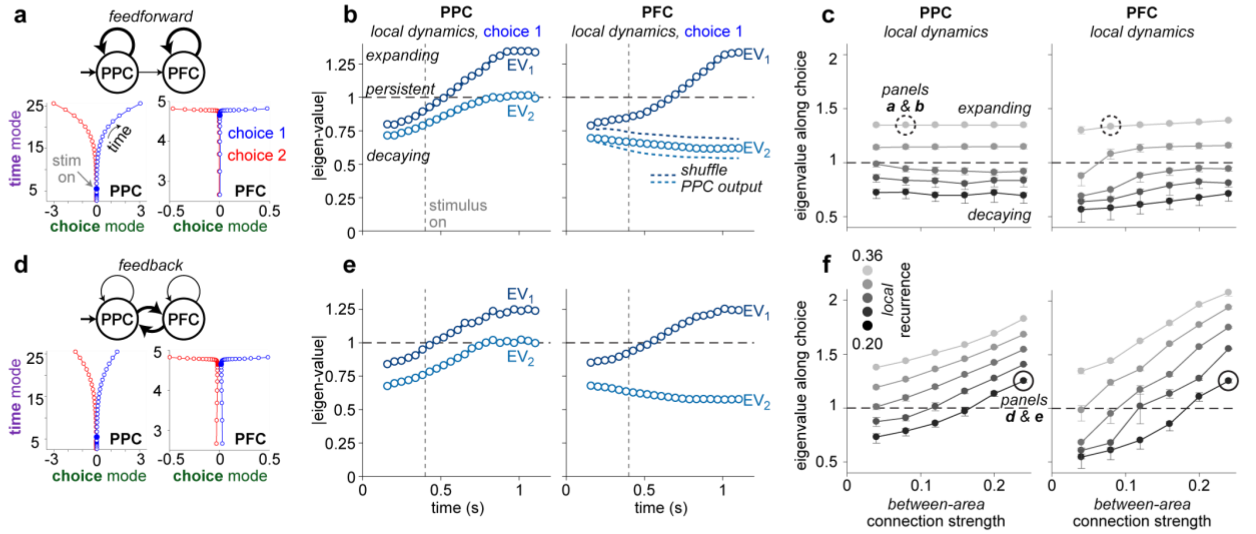


***Figure 6. Local residual dynamics in multi-area networks of perceptual decision making.***

*Each network consists of two interconnected modules (PPC and PFC), whereby a module mimics an RNN with a given strength of local recurrence. PPC is driven by an external input, and feedback connections from PFC to PPC are either absent (**a-c**) or present (**d-f**). **a**, Connectivity (top) and average trajectories (bottom) for an example network with weak feedforward connectivity between areas (top, thin arrow) and strong local recurrence (thick arrows). Condition-averaged trajectories are shown separately for each area for two choices (blue: choice 1, red: choice 2). Trajectories are visualized in a subspace spanned by the choice mode, explaining variance due to choice, and a time mode, explaining condition-independent variance. **b**, Time-varying EV magnitude of the local residual dynamics estimated from residuals in PPC (left) or PFC (right) for choice 1, in the example network in **a**. The external input is turned on 400ms after the start of the trial (gray dashed line). EV magnitudes in PFC are strongly reduced upon shuffling the feedforward output of PPC across trials (blue dashed curves). **c**, Maximum EV magnitude (circle) measured across time for residuals projected onto the choice modes in PPC (left) or PFC (right), as a function of the strengths of local recurrence (black to gray: small to large recurrence) and between-area connections (x-axis). Errorbars indicate 95 percentile bootstrap confidence intervals obtained by fitting residual dynamics to randomly resampled trials (n = 1000). The dashed circle marks the example network shown in **a-b**. **d-f**, Same conventions as in **a-c**, but for networks with between-area feedback.*

The simulations also reiterate the finding that residual dynamics can reflect recurrent computations occurring in an upstream area (Fig. 1b). In the example network with feedforward connectivity, we simulated PFC responses after “shuffling” the output of PPC to remove any temporal correlations ($\boldsymbol{\xi}_t = 0$ in Fig. 1b), while retaining its time-varying mean. In this setting, the EV estimated in PFC fall below 1 (Fig. 6b, PFC; dashed), indicating that local recurrent dynamics in PFC ($\mathbf{F}_{\text{local}}$, Fig. 1b) is actually decaying in these networks. We refer to this effect as an “inflation” of the EV in PFC, due to the correlated input from PPC (Extended Data Fig. 4,5).

Local and long-range recurrent contributions can, however, be resolved by the *global* residual dynamics, estimated from the concurrent, pooled responses from PPC and PFC. We compared global residual dynamics for the two example networks in Fig. 6a,d at the level of the inferred EV (Fig. 7a) and the corresponding eigenvectors (Fig. 7b). The EV magnitudes cannot distinguish between the two networks, with one EV unstable (EV>1), one persistent (EV≈1), and the others

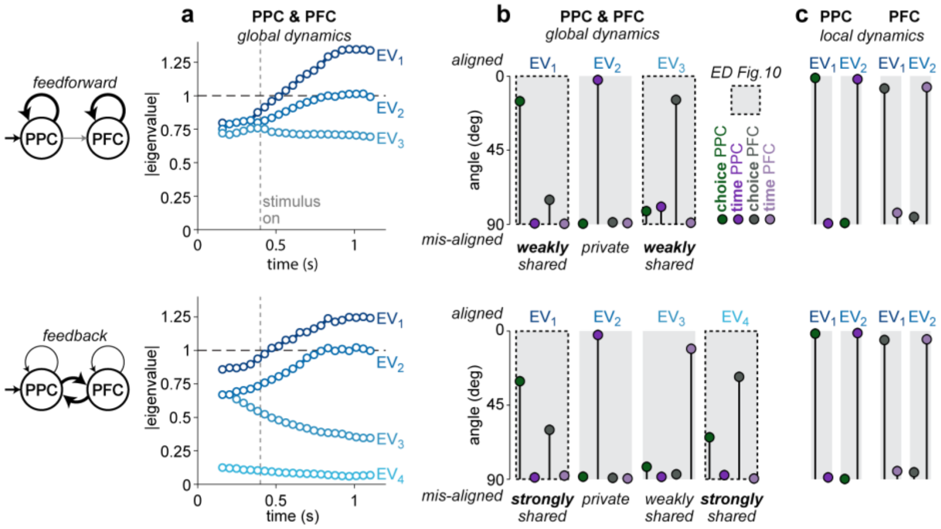


***Figure 7. Global residual dynamics resolves local and long-range recurrent contributions.***

***a**, Time-varying EV magnitudes of the global residual dynamics for the example networks in Fig. 6a (top) and Fig. 6d (bottom). Global residuals are obtained by pooling observations from both areas for a single choice condition (here choice 1). The EV magnitudes do not reliably distinguish between the two example networks. **b**, Alignment (i.e. angle) between the eigenvectors of the global residual dynamics and the choice and time modes in PPC and PFC for the feedforward (top) and feedback (bottom) networks (see legend). Eigenvectors are estimated 0.7s after stimulus onset (dashed line in **a**). Shared eigenvectors span an angle < 90deg with at least one mode in each area. Private eigenvectors are strongly aligned with modes that all lie in a single area. The eigenvector alignments distinguish between the two example networks (top vs. bottom). In particular, the eigenvector aligned with the largest EV (EV$_1$) has a large projection (small angle) onto both the PPC and PFC choice mode in the feedback model (bottom), but only onto the PPC choice mode in the feedforward model (top). **c**, Analogous to **b**, but for the eigenvectors of the local residual dynamics (see Fig. 6b,e) estimated separately based on PPC or PFC responses. The alignment of local eigenvectors does not distinguish between the example networks (top vs. bottom).*

decaying (EV<1; Fig. 7a) in both networks. The number of global EV does not robustly distinguish between networks, as it reflects a somewhat arbitrary cutoff in the dimensions to include in the dynamics subspace (excluded dimensions effectively have EV=0). The eigenvectors of the global residual dynamics, instead, distinguish the two networks. Eigenvectors can be qualitatively categorized as being "shared" across areas, or "private" to an area, depending on whether they have substantial projections (i.e. angle<90) onto choice and times modes (Fig. 7b) in both areas (shared) or only a single area (private). Both networks result in two eigenvectors that are at least partially shared with the choice modes in the two areas, but the relative projections onto each area varies across networks—the two eigenvectors are only "weakly" shared across areas in the feedforward network, whereas they are more "strongly" shared in the feedback network (Fig. 7b; top vs bottom). Notably, these differences are not reflected in the eigenvectors of the local residual dynamics (Fig. 7c, top vs bottom).

## Validating residual dynamics with causal perturbations

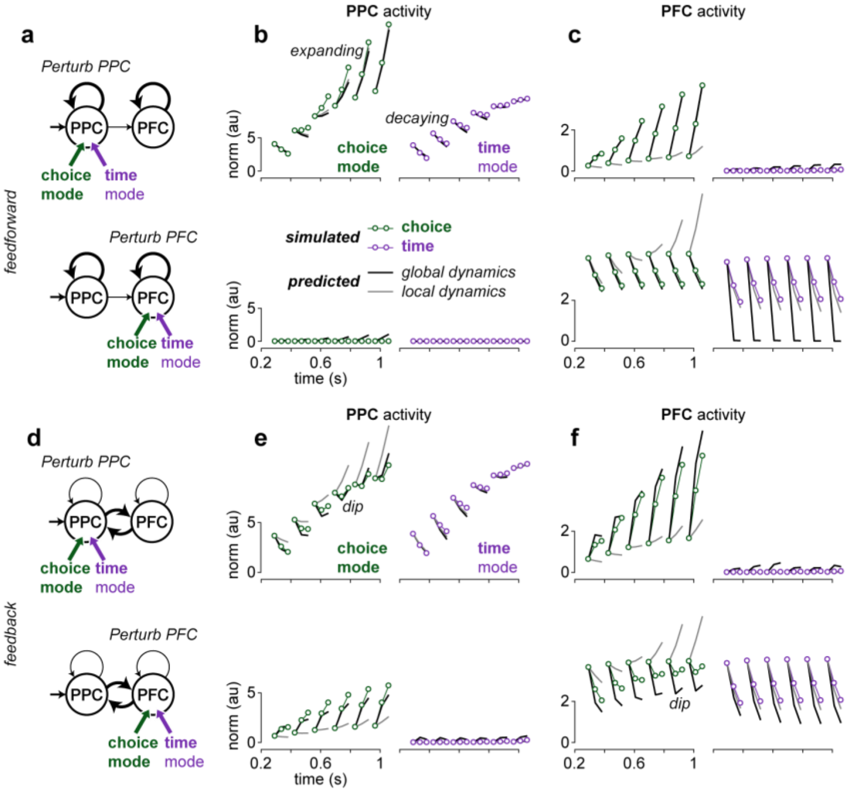


*Figure 8. **Residual dynamics explains the effects of targeted causal perturbations.***

*Simulated responses to brief perturbations for the two example networks in Fig. 6,7 (small circles) are compared to predictions based on residual dynamics (**a-c** and **d-f**: network without and with feedback between areas). Perturbations are applied locally in each area, along the choice or time mode (green and purple circles) at one of six times in the trial (the first point of each curve in **b-c** and **e-f**). Predictions are based either on the local residual dynamics in the simulated area (gray curves; b,e: PPC; c,f: PFC) or on the global residual dynamics (black curves). **a**, Schematic of the location and type of perturbations shown in **b** and **c** for the network without feedback. **b**, Simulated responses in PPC for perturbations in PPC (top) or PFC (bottom) along the respective choice (left) and time modes (right) compared to the corresponding predictions based on local PPC residual dynamics (gray) or global residual dynamics (black). The norm of the population response (y-axis) is shown against time in the trial (x-axis). The last two points on each curve correspond to responses for the two time-steps following the offset of each perturbation. **c**, Analogous to **b**, but for responses in PFC. **d-f**, Analogous to **a-c**, but for the network with feedback. Predictions based on the global, but not the local, residual dynamics capture the qualitative features of the simulated responses to perturbations- that is, decay (**c**, bottom left), expansion (**c**, top left), and dip (decay followed by expansion; **e**, top-left).*

Estimates of residual dynamics, which describe the evolution of "natural" perturbations (Extended Data Fig. 1a), provide predictions of the consequences of "causal" perturbations of the recorded neural population[18,26–29]. We illustrate such predictions for local perturbations applied to PPC or PFC in the example two-area networks (Fig. 8). We simulated perturbations by "injecting" an activity pattern corresponding to the choice mode or the time mode in one area. We applied a brief perturbation at one of six different times after stimulus onset and let the activity evolve under the influence of the recurrent dynamics and the input. The effect of a given perturbation is summarized as the time-varying norm of the population activity in PPC and PFC for a brief time-window following the onset of the perturbation, averaged over many trials (Fig. 8b-c,e-f; a group of three connected points). We compared these simulated perturbations (Fig. 8, dots) to predictions based on the inferred global and local residual dynamics (Fig. 8, black and gray curves).

The effects of perturbations depend on the area where they are applied (Fig. 8, top vs. bottom row in each panel), the perturbed mode (Fig. 8; green: choice, purple: time), and the time within the trial (Fig. 8b-c, e-f, x-axis), and vary across the two example networks (Fig. 8, a-c vs. d-f). Depending on these factors, activity after a perturbation can be expanding, decaying, or show a brief dip (Fig. 8, see labeled examples). This simulated activity is mostly captured, at least qualitatively, by the global predictions (Fig. 8, dots vs. black curves). Qualitative mistakes in the global predictions occur primarily for components of the activity that are very small, like activity in PPC in the feedforward network after a PFC perturbation (Fig. 8b, bottom). Overall, the local predictions fare worse (Fig. 8b-c,e-f; global: black, $R^2 = 0.97$; local: gray, $R^2 = 0.93$). For example, the decay following perturbations of the PFC choice mode in the feedforward network are captured by the global prediction, but not the local prediction (Fig. 8c, bottom-left). The erroneous local prediction is expanding at late trial times, a reflection of the inflation of local EV in PFC in this network (Fig. 6b, PFC; dots vs. dashed). In the feedback network, PPC and PFC perturbations along the choice mode lead to a dip in activity in the perturbed area (Fig. 8e, top-left and Fig. 8f, bottom-left) and to expanding activity in the non-perturbed area (Fig. 8f, top-left and Fig. 8e, bottom-left). These dependencies are qualitatively captured by the global predictions, but not the local predictions. The observed dips reflect the existence of a global, shared unstable direction, which local residual dynamics cannot adequately capture (Extended Data Fig. 10).

## Discussion

The properties of residual responses provide insights into the nature of recurrent computations underlying neural population dynamics. Our analysis of residual dynamics extends previous work that leveraged trial-by-trial variability to understand neural computations[20,21,23,24,44], by providing a full, quantitative description of the time-varying dynamics of population-level trial-by-trial variability. Our approach can capture dynamics that are globally non-linear[9], through a series of local approximations capable of resolving differences in dynamics across state-space locations and time.

Response residuals are computed by discounting the component of neural responses that is repeatable across trials of a given task condition, and can therefore be explained with more easily interpretable models than previous descriptions of the full single-trial neural response[5–7].

Discounting this component does not necessarily remove all sources of external inputs into the recorded area (Fig. 1a), implying that residual dynamics in a single area may not reflect only the local recurrence in the recorded area. Instead, residual dynamics reflects the combined effects of local recurrence and recurrent dynamics unfolding within the output space of upstream areas that provide an input to the recorded area (Fig. 1b, Fig. 6, Extended Data Fig. 4).

The contributions from local and long-range recurrence to neural responses can be distinguished by inferring the global residual dynamics, based on recordings from the entire network of inter-connected areas (Fig. 7). The resulting description of dynamics in terms of modes (i.e., eigenvectors) that are shared across areas[41], or private to a single area, relates to previously identified communication- and null-subspaces between areas[25,44,52,53]. Global residual dynamics goes beyond a static description of such subspaces, as it captures also the dynamics of responses resulting from unidirectional or bidirectional communication between areas. In particular, global residual dynamics leads to fine-grained predictions of the consequences of small causal perturbations that probe the intrinsic manifold explored by the neural variability[29,30] (Fig. 8).

Our local estimates of PFC residual dynamics provide constraints on the properties of recurrent dynamics implemented by the recorded PFC population and its contributions to decision-making and movement generation. The largest estimated time constants provide an upper bound on the time-constants of the local recurrent dynamics in PFC (Fig. 4e; 322ms and 503ms in monkeys T and V; medians, n = 352: 2 choices x 4 configurations x 44 times in trial), as any upstream contribution to PFC responses would typically inflate these estimates (Fig. 6b; Extended Data Fig. 4,5). Recurrent dynamics in PFC is thus slow[54,55], but stable throughout the decision and movement epochs. This finding does not rule out that the decision process leading to the monkeys' choices involves unstable or line-attractor dynamics (Fig. 1c), but those dynamics would have to unfold in areas upstream of PFC[56], and at least partly outside their communication subspace with PFC.

The estimated time-constants would reflect the dynamics of the decision process if that process unfolded either in PFC alone, or within its communication subspace with other areas (as for all networks in Fig. 6). In such scenarios, our estimates imply leaky evidence accumulation (Fig. 1c, point attractor), whereby late evidence affects choice more strongly than early evidence. In practice though, monkeys often terminate evidence accumulation early in the trial, when a decision threshold is reached[57], which would reduce the behavioral effects of leaks in the accumulation. Notably, a recent study hypothesized that the termination of evidence accumulation coincides with the onset of rotational dynamics in PFC[58]. In our study, condition-independent, rotational dynamics during the decision epoch also stands out, as in monkey T it is the component of the recorded activity that can be best explained as resulting from recurrent computations (Fig. 5). Irrespective of the possible contributions of PFC to the process underlying the monkeys' choices, this finding may be indicative of a broader role for PFC in governing transitions between cognitive states[58,59], e.g. the transition from an uncommitted to a committed state.

Around the time of the saccade, PFC residual dynamics is quickly decaying, largely non-rotational, and only weakly non-normal, implying that PFC does not implement rotational dynamics[13,36], dynamic attractors[37], or strongly non-normal[60] recurrent dynamics of the kind previously proposed to explain movement activity in motor cortex. Rotational dynamics and dynamic attractors are also unlikely to be implemented in an upstream area driving PFC movement responses through a communication subspace, since the signatures of those dynamics would then also appear in PFC residuals (Fig. 6, Extended Data Fig. 4). Strong non-normal dynamics in an upstream area, however, could possibly explain the observed PFC responses. Non-normal systems can generate large activity transients that project only weakly onto the activity subspace containing the slowest dynamics. If the output from such an upstream area was partially aligned with the activity transients, but orthogonal to the slow dynamics, it could possibly drive strong "input-driven" movement-related activity in PFC without revealing the signatures of the strongly non-normal dynamics that created it. Alternatively, the mismatch between average trajectories and residuals in the movement epoch could reflect a failure in our estimation procedure. For one, estimates of residual dynamics become biased when trial-by-trial variability is too small, which however does not seem to be the case in our data (Extended Data Fig. 9). For another, dynamics during movement may be strongly non-linear, and thus not well approximated by our local linear description (Extended Data Fig. 1). In both scenarios, our estimated dynamics would not provide a good description of the true dynamics.

Finally, residual dynamics may provide insights into more general biological constraints at play in the underlying neural circuits. The inferred EV are smaller than but close to 1 during the decision epoch, consistent with circuits operating near a critical regime, resulting in large variability and sensitivity to inputs[40,61–63]. Single-neuron variability is transiently reduced at the time of stimulus and movement onset (Extended Data Fig. 7), potentially reflecting the widespread quenching of variability in response to task events[21,64]. Near-critical dynamics, non-normality, and variability quenching emerge naturally in balanced excitation-inhibition (E-I) networks[65,66]. A disruption of E-I balance by the onset of an input could lead to contracting dynamics and reduced variability. In our PFC recordings, reduced variability coincides with contracting dynamics at movement onset, but not at stimulus onset (Extended Data Fig. 7). This finding suggests that current models of E-I networks[65,66] may have to be adapted to fully capture the interactions of internal dynamics, inputs, and variability we observed in PFC.

## Acknowledgements

We thank John Reppas and William Newsome for the data collection. We thank Kevan Martin and all members of the Mante Lab for their valuable feedback, as well as Nicolas Meirhaeghe, Lea Duncker and Mehrdad Jazayeri for discussions and comments on the manuscript. This work was funded by Swiss National Science Foundation (Award PP00P3-157539, VM), the Simons Foundation (SCGB 328189 and 543013, VM; SCGB 543039 and 323228, MS), the Swiss Primate Competence Center in Research (VM), the Gatsby Charitable Foundation (MS), the Howard Hughes Medical Institute (William Newsome), and the Air Force Research Laboratory (William Newsome)

## Author Contributions

A.R.G and V.M conceived and designed the study. A.R.G developed the methods and performed the analyses, with input from M.S. and V.M. A.R.G and V.M wrote the manuscript. All authors were involved in discussing the results and the manuscript.

## Competing Interests Statement

The authors have no competing interests to disclose.

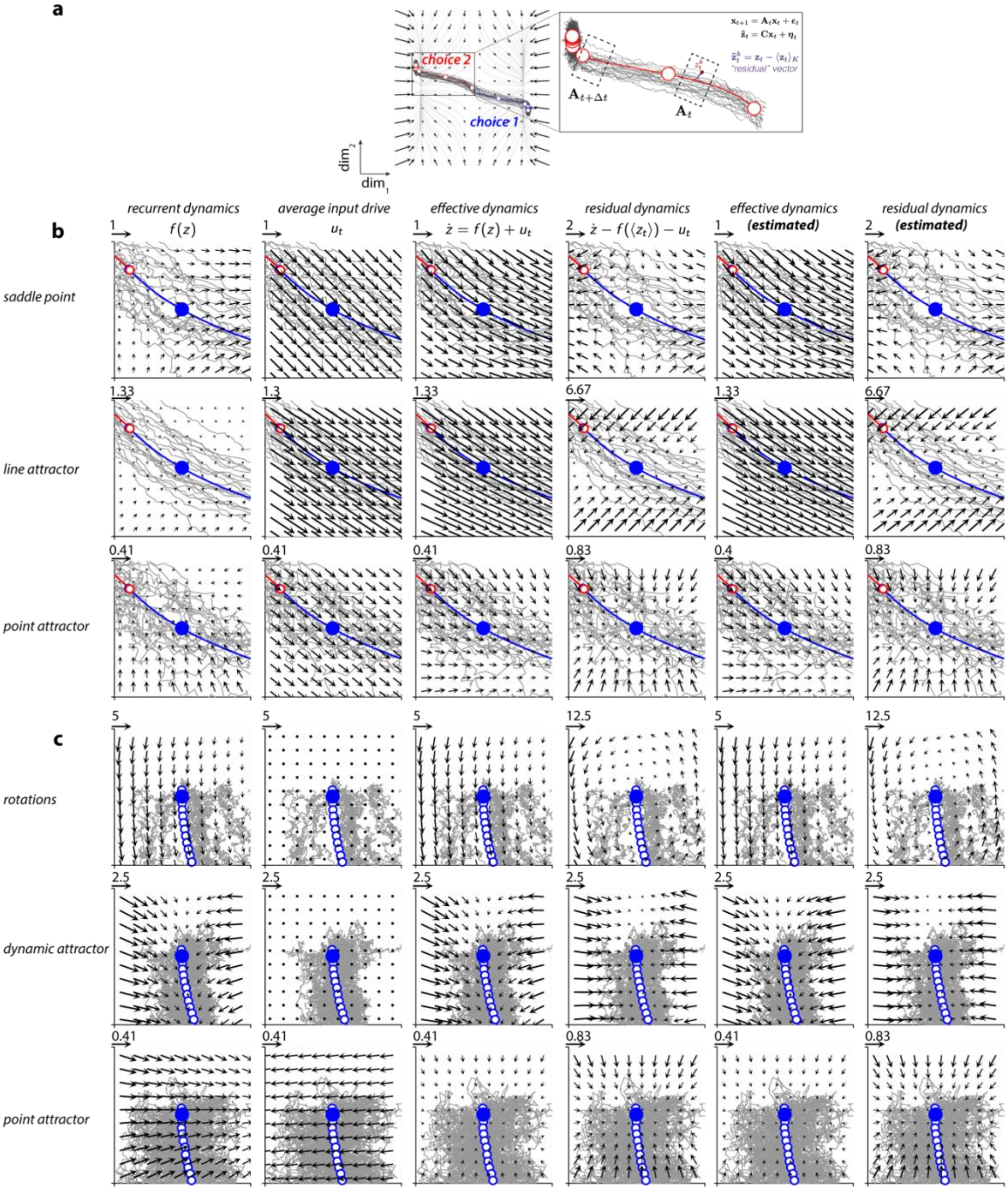
## Data Availability

All neural data used in the manuscript are available at *https://doi.org/10.5281/zenodo.7378387*

## Code Availability

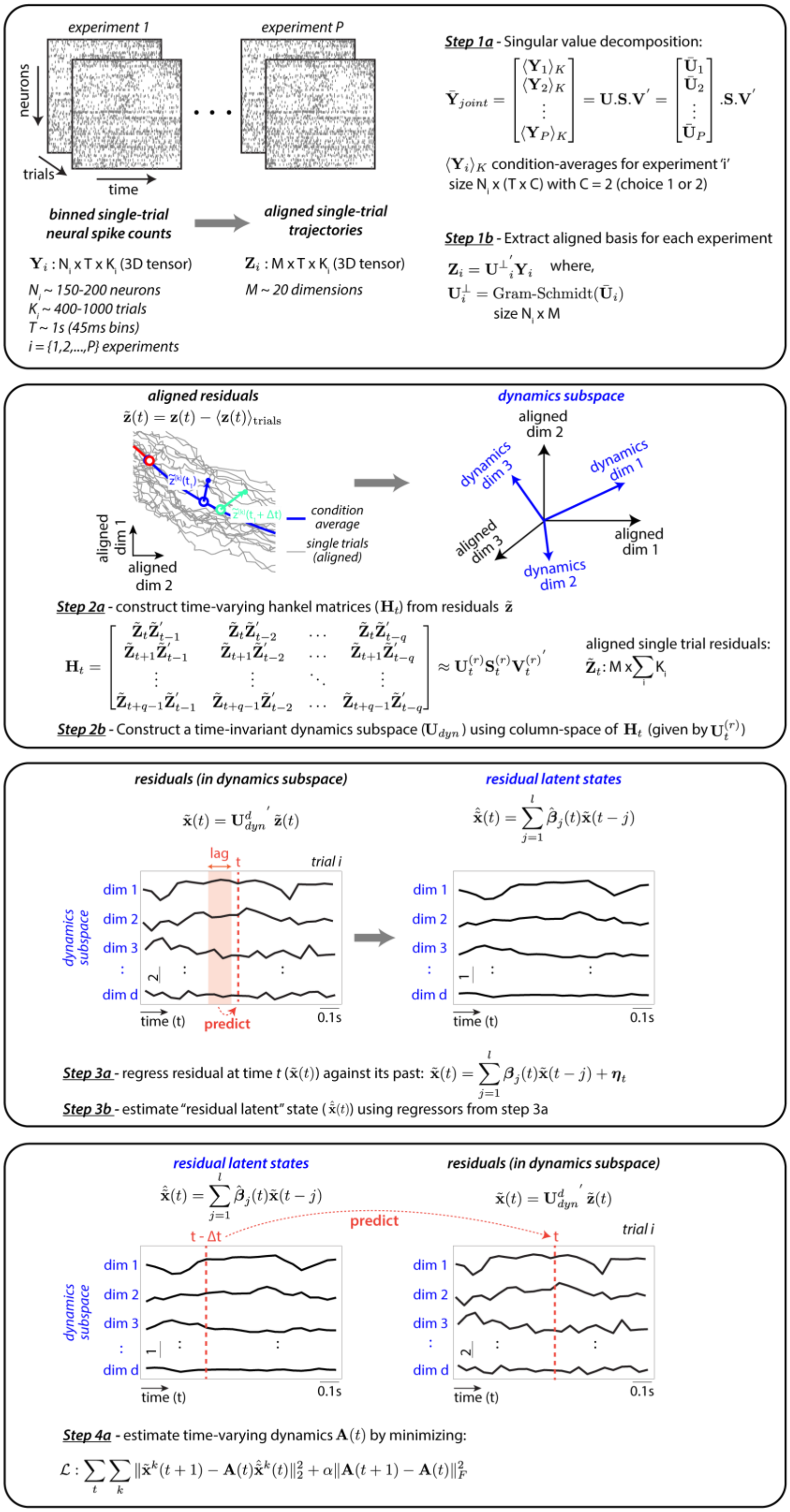
The data analysis pipeline and code to generate simulations presented in the paper are available at *https://github.com/anirgalgali/residual-dynamics*

# Extended Data Figures



***Extended Data Fig. 1: Residual and effective dynamics in models of decisions and movement***

***a**, Variability in responses across trials from the same task condition are interpreted as perturbations away from the condition-averaged trajectory. The evolution of these perturbations reflects the properties of the underlying recurrent dynamics (flow field, same conventions as in Fig. 1c). Inset on right shows a magnified view of the condition-averaged trajectory (red, choice 2) and corresponding single trials (dark gray) simulated from the saddle point model. Residual vectors at each time (shown in purple for a single trial and time) are computed by subtracting the condition-averaged response at that time from the corresponding single-trial response (purple equation). Time-varying dynamics matrices (**A**$_t$) of a linear time-varying, autonomous state-space model* *(black equations, top-right) are fit to the residuals. These matrices approximate the dynamics in distinct 'local' regions of state space (e.g. dashed boxes) and are indexed according to time and condition.* ***b-c****, Components of the dynamics for the models of decisions (****b****) and movement (****c****) for an example reference time (blue dot) along the condition-averaged trajectory for choice 1. Same conventions as in Fig. 2a. Dynamics are shown for a local state-space region close to the corresponding initial condition (boxes in Fig. 1c, d; left). For all models, the estimated effective and residual dynamics (columns 5 and 6) closely match the true effective and residual dynamics (columns 3 and 4). In these models, the residual dynamics (column 4) reflects only the recurrent dynamics (column 1), but is not identical to it. For one, the fixed point of the residual dynamics by definition is located at the location of the reference state (the blue dot), which in general does not match the position of fixed points of the recurrent dynamics (e.g. the red circle in the first row and first column, corresponding to the position of the unstable fixed point in the saddle point model). The position of fixed points of the recurrent dynamics can only be inferred if the inputs are known, a requirement that is not fulfilled in many experimental settings. For another, consistent drifts resulting from the recurrent dynamics (e.g. the drift along the channel in the dynamic attractor model) are not reflected in the residual dynamics. Such drifts are "subtracted" from the variability in the computation of residuals. Differences in the underlying recurrent dynamics are more apparent in the residual compared to the effective dynamics in cases where the input drive is strong. For example, the average cosine similarity between flow fields is 0.27/0.99 (saddle vs. line-attractor), 0.02/0.94 (saddle vs point-attractor) and 0.58/0.95 (line-attractor vs point-attractor) for the residual/effective dynamics.*



### *Step 1: Session alignment*

**Goal:** identify common, aligned modes across different experiments.

**Approach:** singular value decomposition and Gram-Schimdt orthogonalization

**Input:** concatenated (across sessions) condition averaged trajectories $\bar{\mathbf{Y}}_{joint}$

**Output:** "aligned" single-trial trajectories

**Hyperparamters:** # of aligned modes (M)

### *Step 2: Dynamics subspace estimation*

**Goal:** identify dimensions that are predictive of the "future" based on the "past".

**Approach:** Hankel matrix decomposition

**Input:** "aligned" residual responses

**Output:** ordered set of dimensions that constitute a "dynamics subspace"

**Hyperparamters:** rank of hankel matrix (r) (ascertained using cross-validation)

### *Step 3: Residual latent state estimation*

**Goal:** estimate the residual latent state

**Approach:** First stage of a two-stage least squares(2SLS) regression approach

**Input:** "aligned" residual responses & dimensions of dynamics subspace (step 2)

**Output:** denoised estimate of residual latent state at each 't'

**Hyperparamters:** dimensionality of dynamics subspace (d) & number of regression lags (l)

### *Step 4: Time-varying dynamics estimation*

**Goal:** estimate time-varying residual dynamics

**Approach:** Second stage of a two-stage least squares(2SLS) regression (with regularization)

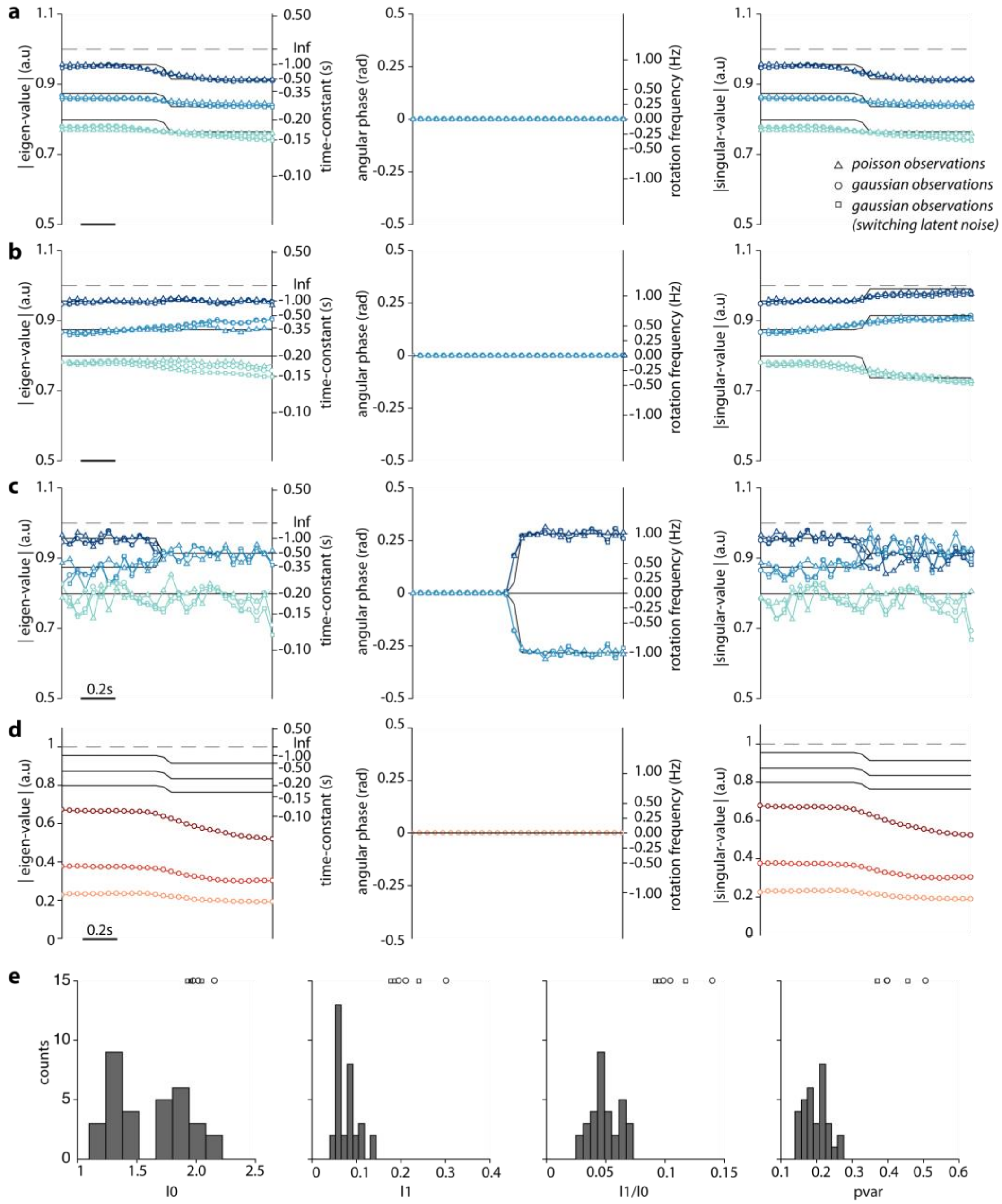
**Input:** "aligned" residual responses & first stage estimates of residual latent states (from step 3)

**Output:** time-varying dynamics matrices $\mathbf{A}_t$

**Hyperparamters:** smoothness penalty α (ascertained using cross-validation)

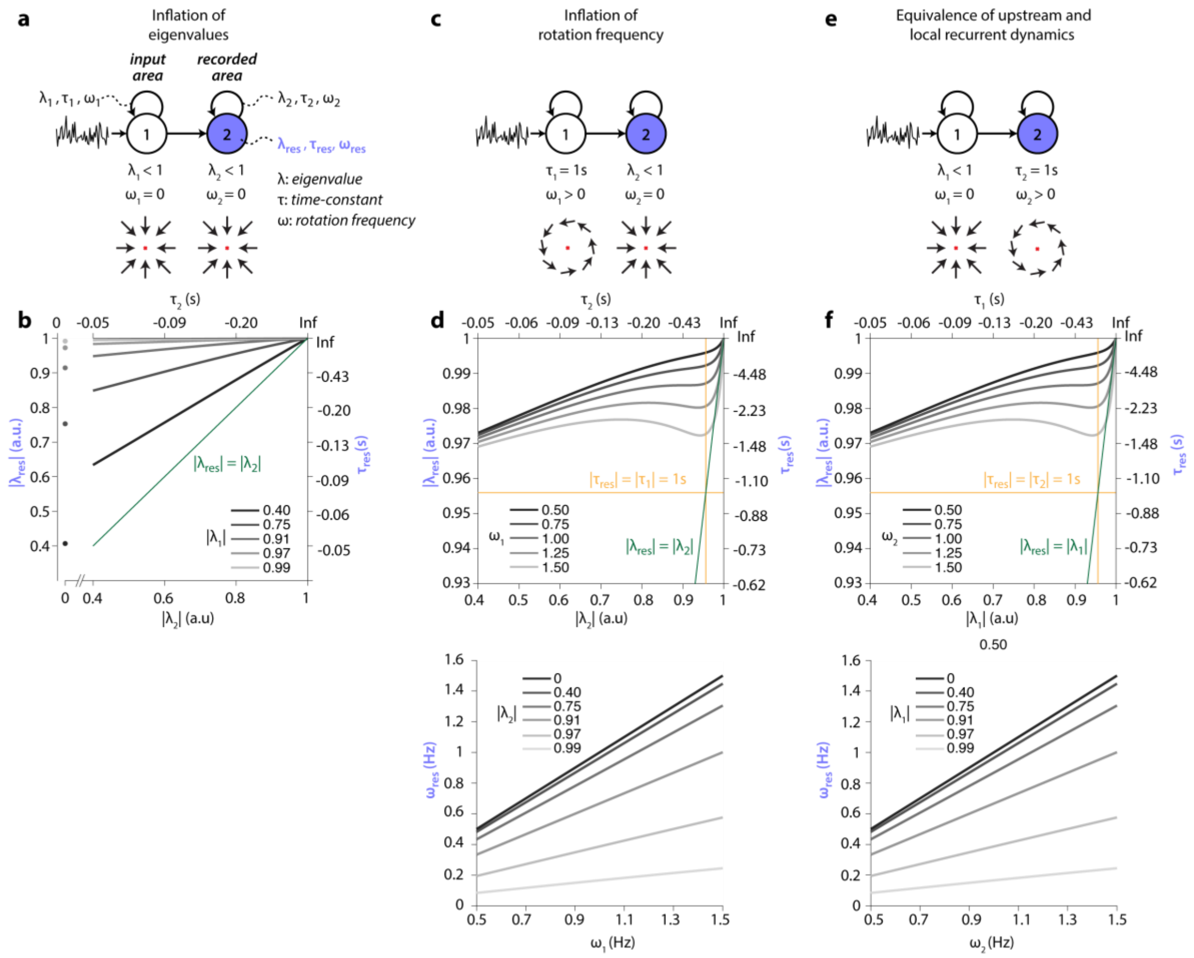
*Extended Data Fig. 2: Schematic of analysis pipeline*

*Schematic depicting the complete data analysis pipeline for inferring residual dynamics from noisy neural population recordings (see Methods). The pipeline involves four sequential steps. Step 1: session alignment; involves pooling single trials from different recording sessions to increase the statistical power of the analyses. Step 2: dynamics subspace estimation; involves using 'aligned' single-trial neural residuals to obtain estimates of a dynamics subspace (*$\boldsymbol{U}_{dyn}$*) that effectively contains the residual dynamics. Step 3: residual latent state estimation; involves using the first stage of a two stage least squares (2SLS) approach to estimate a 'denoised' latent residual state. Step 4: time-varying dynamics estimation; uses the denoised residual latent states (obtained in step 3) for the second stage of the 2SLS, to estimate the time-varying residual dynamics matrices (*$\boldsymbol{A}_t$*).*



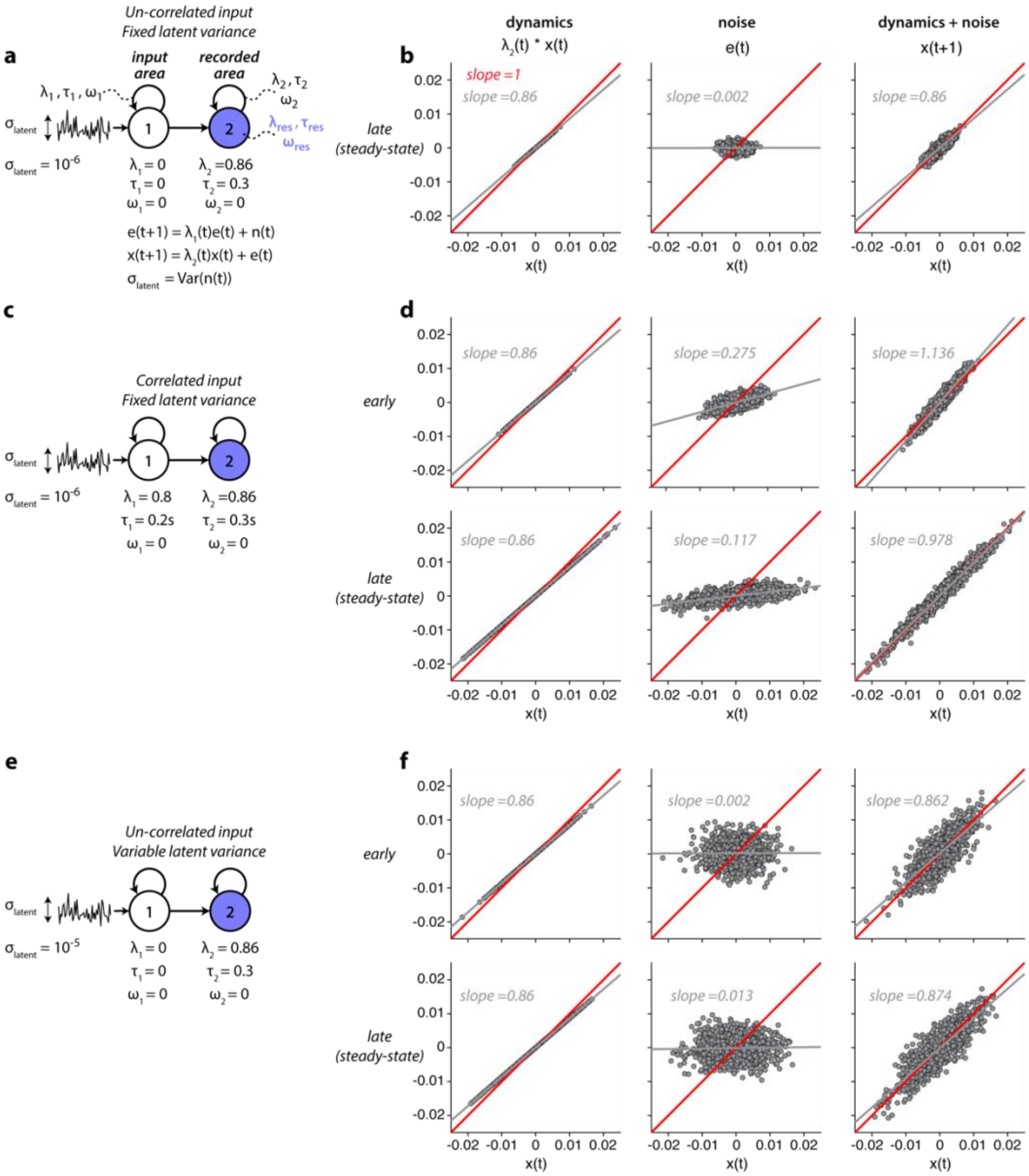
***Extended Data Fig. 3: Residual dynamics of simulated, time-varying, linear dynamical systems.***

***a-c**, Validation of the estimation procedure on simulations of time-varying, linear dynamical systems with temporally uncorrelated latent noise (see Methods; Supplementary Methods). Simulations are based on a latent variable dynamical system with 3 latent dimensions and 20 observed dimensions. Residual responses are generated using a gaussian (circle markers: fixed latent noise variance; square markers: latent noise variance switches mid-way through the trial) or poisson (triangle markers) observation process. In all simulations, the properties of the dynamics switch midway through the simulated time window, from slowly decaying to quickly decaying (**a**); from normal to non-normal (**b**); or from non-rotational to rotational (**c**). As in Fig. 4b-d, we characterize dynamics with the magnitude of the eigenvalues (left), the rotational frequency (middle), and the singular values (right). Markers correspond to the estimated residual dynamics, black curves to the ground-truth values. The estimated residual dynamics accurately matches the ground-truth for all types of dynamics and observation models, before and after the switch, and also reveals the time of the switch. We observed this match even when the latent noise variance of gaussian observations was switched at the same time as the eigenvalues/eigenvectors of the dynamics (square markers), demonstrating that estimates of* *residual dynamics are robust to changes in latent noise variance (see also Extended Data Fig. 5a-b vs e-f).* ***d****, Analogous to* ***c****, but for residual dynamics (circles) estimated using ordinary least squares (OLS) instead of two-stage least squares (2SLS) as in* ***c****. Results are only shown for data simulated using a gaussian observation process. Unlike the 2SLS estimates, the OLS estimates are strongly biased, i.e. the magnitude of the eigenvalues and the singular values are consistently underestimated. These biases are expected—they arise because both the regressors and the dependent variables are corrupted by observation noise (see Methods). The 2SLS instead produces unbiased estimates, as the first stage of 2SLS results in a denoising of the regressors (Methods; see also Extended Data Fig. 9).* ***e****, Parameters of the latent noise and observation noise for the simulations in* ***a-d*** *were chosen to approximately match the variability in the measured PFC responses. The variability in the measured responses were quantified in terms of four statistics (l0, l1, l1/l0 and pvar, x-axis; see Supplementary Methods). Histograms indicate the respective values of these statistics in the neural data (one data point per task configuration, choice condition and monkey; see legend in Extended Data Fig. 6a). The open markers (top, same conventions as* ***a-c****) indicate the values of the statistics in the simulations for each of the three models.*



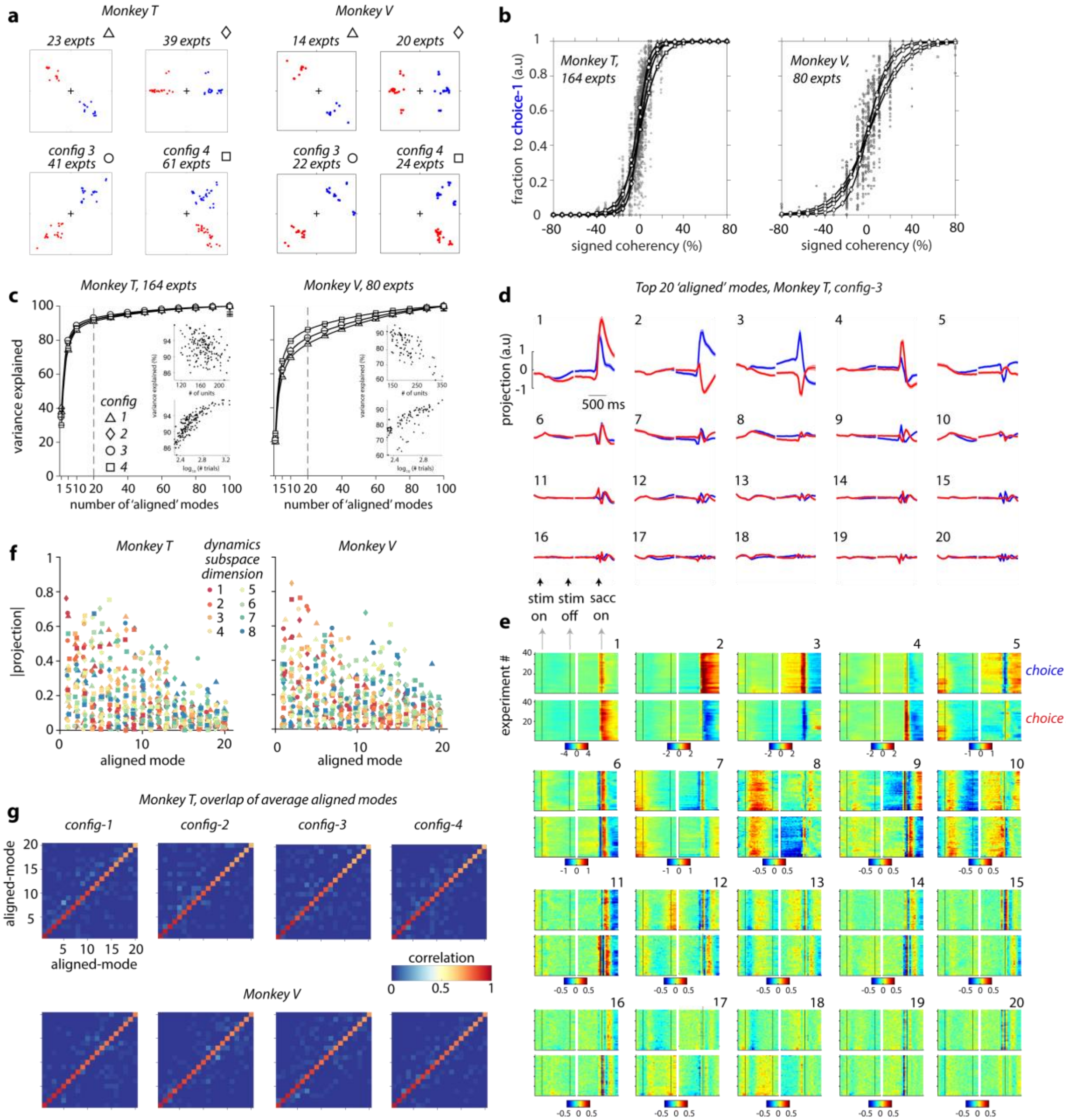
**_Extended Data Fig. 4: Inflation of local residual dynamics in a linear two-area dynamical system_**

*We systematically explored the effect of correlated input variability on estimates of residual dynamics in a two-area, linear dynamical system (see Methods & Supplementary Methods). The input area implements 2D isotropic recurrent dynamics characterized by parameters* $\boldsymbol{\lambda_1}$, $\boldsymbol{\tau_1}$, *and* $\boldsymbol{\omega_1}$ *(eigenvalue, time-constant, rotation frequency). Activity in the input area is externally driven by uncorrelated noise. Values of* $\boldsymbol{\lambda_1}$ *closer to 1 result in longer auto-correlation times in the variability of activity in the input area. This activity provides the input into the recorded area, which implements 2d isotropic recurrent dynamics with parameters* $\boldsymbol{\lambda_2}$, $\boldsymbol{\tau_2}$, $\boldsymbol{\omega_2}$. *Residual dynamics at steady-state is estimated from activity of the recorded area. At steady state, estimates can be derived analytically (see Supplementary Math Note B). Because of temporally correlated input variability, the properties of the residual dynamics (*$\boldsymbol{\lambda_{res}}$, $\boldsymbol{\tau_{res}}$, $\boldsymbol{\omega_{res}}$*) in general do not match those of the recurrent dynamics in the recorded area.* ***a-b****, Inflation of eigenvalues.* ***a****, Schematic of the model (top) and recurrent dynamics in each area (bottom, flow fields). Recurrent dynamics is stable and non-rotational in both areas.* ***b****, Residual dynamics (*$\boldsymbol{\lambda_{res}}$*) in the recorded area as a function of recurrent dynamics in the recorded area (*$\boldsymbol{\lambda_2}$*, x-axis) and in the input area (*$\boldsymbol{\lambda_1}$*, gray lines). The eigenvalues of the residual dynamics are inflated, i.e.* $\boldsymbol{\lambda_{res}}$ *is larger than* $\boldsymbol{\lambda_2}$ *(all gray lines above the green line). Larger* $\boldsymbol{\lambda_1}$ *(longer input auto-correlations) lead to stronger inflation. For* $\boldsymbol{\lambda_2}$ *= 0 (no recurrent dynamics in the recorded area)* $\boldsymbol{\lambda_{res}} = \boldsymbol{\lambda_1}$ *(gray circles).* ***c-d****, Inflation of rotation frequency.* ***c****, Recurrent dynamics is rotational in the input area, but stable and non-rotational in the recorded area.* ***d****, Residual dynamics in the recorded area, expressed as the magnitude of the eigenvalue (*$\boldsymbol{\lambda_{res}}$*, top) and the rotation frequency (*$\boldsymbol{\omega_{res}}$*, bottom). The eigenvalues of the residual dynamics are generally inflated (top), but the relation with* $\boldsymbol{\lambda_2}$ *is non-monotonic and depends on* $\boldsymbol{\omega_1}$. *The residual dynamics is rotational (bottom,* $\boldsymbol{\omega_{res}} > \mathbf{0}$*) even though the recurrent dynamics in the recorded area is not (*$\boldsymbol{\omega_2}$*= 0). The inflation of rotation frequency is reduced for increasing* $\boldsymbol{\lambda_2}$. ***e-f****, Equivalence of upstream and local recurrent dynamics.* ***e****, Analogous to* ***c****, but dynamics is switched between input and recorded area.* ***f****, Analogous to* ***d****, but for the dynamics in* ***e****. The residual dynamics is identical to that in* ***d****. In general, residual dynamics in the recorded area reflects the combined effect of local and upstream recurrent dynamics.*



***Extended Data Fig. 5: Explanation of input driven inflation in residual dynamics***

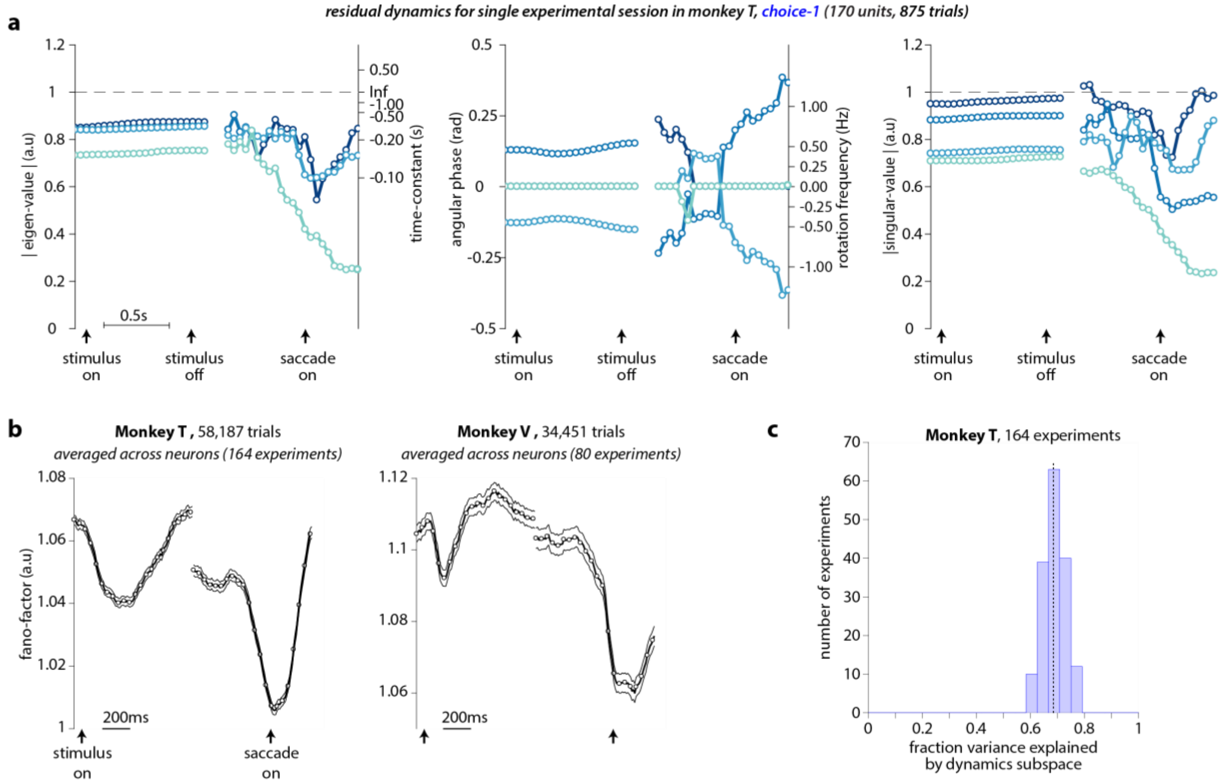
*To gain an intuitive understanding of inflation of eigenvalue magnitude, we consider simulations of two-area linear dynamical systems similar to those in Extended Data Fig. 4a. For simplicity, here we simulate stable 1d-dynamics in each area, whereby variability of the input into the recorded area is either temporally correlated (**c-d**) or uncorrelated (**a-b**, **e-f**), and has fixed (**a-b**, **c-d**) or time-dependent latent noise variance (**e-f**). The variability injected into the input area is always temporally uncorrelated. Recurrent dynamics in the recorded area is identical in all simulations. **a**, Model parameters for the case of temporally uncorrelated input ($\lambda_1$ = 0). **b**, Contributions to activity x in the recorded area at steady-state. Activity x(t) (x-axis) is propagated through the recurrent dynamics (left, y-axis) and added to the noise e(t) (middle, y-axis) to obtain activity x(t+1) at time t+1 (right, y-axis). The noise e(t) corresponds to activity/output of the input area, and is shaped by dynamics determined by $\lambda_1$. Points in the scatter plots correspond to different simulated trials. Estimating the eigenvalue of the residual dynamics in the absence of observation noise amounts to measuring the slope of the regression line relating x(t) to x(t+1) (right, gray line). In this case, this slope is identical to that obtained if the latent noise had not been added to the activity (left, gray line), meaning that residual dynamics correctly reflects the effect of the recurrent dynamics in the recorded area (slope < 0, reflecting $\lambda_2$ < 0; left). **c**, Model parameters for the case of correlated input ($\lambda_1$ > 0 for t > 0; $\lambda_1$ = 0 at other times). **d**, Analogous to **b**, but for the model in **c**. Here activity and noise* *are shown at two times in the trial: early, when steady-state is not yet reached (top) and late, at steady-state (bottom). At both times, residual dynamics is inflated, i.e. the regression slope between x(t) and x(t+1) (right) is larger than that obtained by applying only the recurrent dynamics (left), indicating inflation of the eigenvalues. Inflation occurs because the noise itself is correlated with activity in the recorded area (middle, slope > 0), an effect that results indirectly from the correlation between e(t) and e(t-1). At steady state, even the inflated residual dynamics is still stable (bottom-right, slope < 1; see also Extended Data 4b). However, immediately after the onset of the temporally correlated input, residual dynamics erroneously reveals an instability (top-right, slope > 1).* ***e***, *Parameters for the case of temporally uncorrelated noise but time-varying noise variance. The variance of the noise injected into the input area is increased at time t = 0, from* $\sigma_{latent}$ *=* $10^{-6}$ *to* $10^{-5}$. ***f,*** *A change in noise variance does not result in inflation of the residual dynamics, neither early nor late after the change (right, top and bottom; same slope as on the left; see also Extended Data Fig. 3a-c, squares).*



***Extended Data Fig. 6: Alignment of neural population responses from different experiments.***

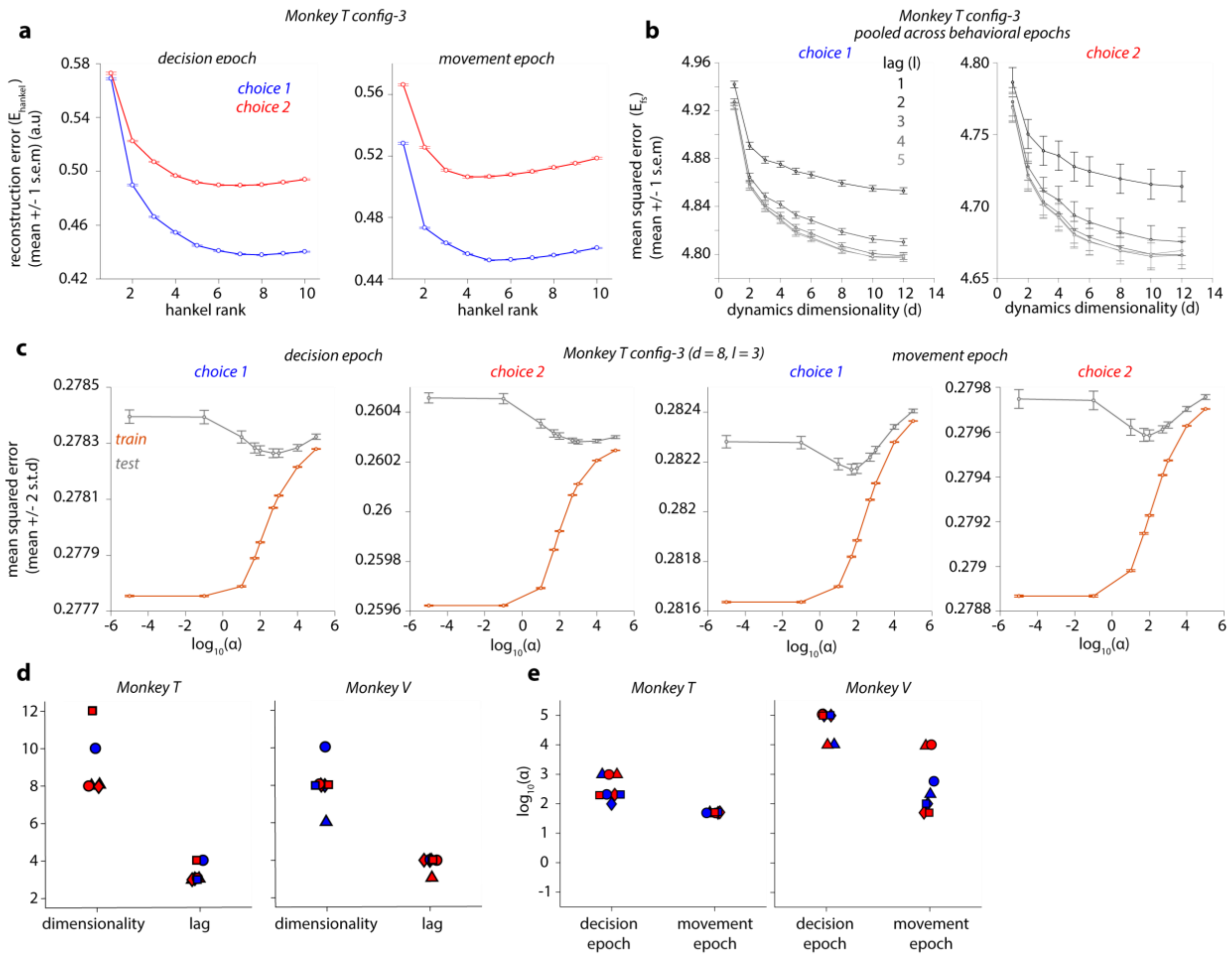
*Validation of the session alignment procedure of the analysis pipeline (Extended Data Fig. 2, Step 1; see Methods). We aligned neural population responses of all experiments belonging to the same task configuration and then pooled the aligned single trial responses across experiments before computing the residuals used in estimating the dynamics. The outcome of the session alignment procedure is a set of 20 'aligned' modes for each experiment, defined such that the activity of each mode has the same dependency on time and choice across experiments.* ***a****, Definition of task configurations. We assigned each experiment to one of four target configurations (distinguished by markers, indicated on top of each panel along with number of experiments) based on the angular position of the targets (blue: choice 1; red: choice 2). The position of the targets was similar, but not identical, across experiments within the same task configuration. (left: Monkey T, right: Monkey V).* ***b****, Psychometric curves for all experiments in both monkeys (left: Monkey T, right: Monkey V), showing the fraction of saccades to choice 1 as a function of the signed motion coherency. Each gray data point is computed from trials belonging to a single experiment. The employed values of signed coherency varied slightly across experiments, in an attempt to achieve a comparable overall performance in each experiment.*

*Black curves show logistic functions fitted separately to data points from a given task configuration (different markers; see legends in c) and evaluated at logarithmically spaced levels of coherency (positions of the white markers along the x-axis).* ***c****, Cumulative variance explained in condition-averaged population responses (mean +/- 2 s.e.m. across experiments; symbols as in* ***a****, n = number of experiments in each task configuration: see* ***a****) as a function of the number of aligned modes in both monkeys (left: Monkey T, right: Monkey V). The cumulative variance explained by the first 20 aligned modes for all 164 experiments in Monkey T and 80 experiments in Monkey V showed a strong positive trend with number of trials (inset, bottom) and a weak negative trend with the number of units (inset, top).* ***d****, Activity of the first 20 aligned modes (numbered from top-left to bottom-right) for config-3 in monkey T (15,524 trials across 41 experiments) ordered according to the amount of variance explained. Activity is defined as the projection of the population condition averages onto each mode. The projection was computed separately across experiments for choice 1 and choice 2 (blue and red) with responses aligned either to stimulus onset or saccade onset (black arrows). The resulting projections were then averaged across experiments (line: mean; shading: 2 s.e.m. across 41 experiments).* ***e****, Same data as in* ***d****, but showing the time-course of each aligned mode (numbered from 1 to 20) for each individual experiment (y-axis) separately for the two choice conditions (choice 1 and choice 2, top and bottom sub-panels). Differences in the activation of a given mode across experiments (i.e. across rows in each sub-panel) are much smaller than the differences in the activations across modes (i.e. across sub-panels), demonstrating the success of the alignment procedure.* ***f****, Absolute value of the projection (y-axis) of the 8 basis vectors (dim-1 through dim-8; red to blue) that span the dynamics subspace ($\boldsymbol{U}_{dyn}$, estimated in Step 2 of the analysis pipeline; Extended Data Fig. 2) onto the 20 aligned modes, indicating the relative alignment of the aligned and dynamics subspace. The dynamics subspace is computed separately for each task configuration (symbols as* ***a****) in each monkey (left: Monkey T, right: Monkey V), and projects most strongly onto the first few aligned components (i.e large projection values for smaller aligned mode number). The dynamics subspace thus largely overlaps with the subspace of activity that captures most of the task-related variance in the responses (see also Extended Data Fig. 7c).* ***g****, Evaluation of the alignment procedure for all task configurations (columns) in both animals (rows). Each element of the matrix is obtained from the correlation coefficient between the time-courses of two aligned modes (i.e. positions along horizontal and vertical axes). We show the median correlation coefficient across all pairs of dissimilar experiments. Values close to 1 along the diagonal and close to 0 in the off-diagonal indicate that the time-courses are much more similar across experiments than across modes, indicating successful alignment.*



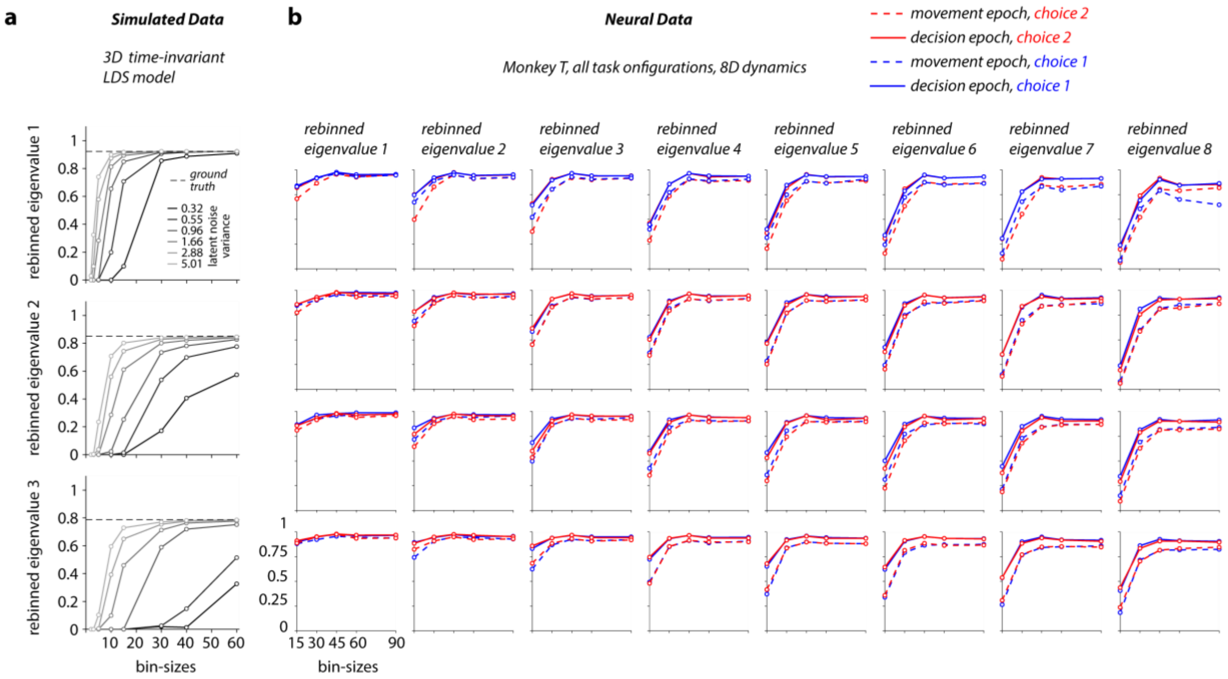
***Extended Data Fig. 7: Single session and single unit results***

***a**, Residual dynamics estimated using neural data for a single choice condition (choice-1, 875 trials) from a single experiment in monkey T. This experiment has the largest number of trials among all experiments in monkey T. Conventions as in Fig 4b-d. We estimated the residual dynamics directly from high-dimensional residual observations that corresponded to square-root transformed, binned spike-count vectors (dimensionality = number of units; 170 for this session), without performing the session alignment (step 1 in Extended Data Fig. 2). Overall, the properties of the residual dynamics estimated from this single session are similar to those obtained after pooling trials across sessions (Fig 4b-d, 8 dimensional), suggesting that the main features of the residual dynamics (Fig. 4) are not affected by the alignment procedure. The lower dimensionality of the estimated residual dynamics (4 dimensions, blue to cyan; compared to 8 dimensions in Fig. 4a-d) most likely is a consequence of the smaller number of available trials in the single session compared to the aligned sessions. The resulting smaller statistical power makes is harder to estimate, in particular, the faster decaying eigenmodes of the dynamics. **b**, Trial-by-trial variability in single neurons is transiently reduced at the onset of specific task-events. We quantified single neuron variability as the time-varying, mean-matched Fano-factor computed by pooling units/neurons across all experiments in a monkey (empty circles: mean; dashed curve: 95% normal confidence intervals obtained by resampling datapoints; left: Monkey T, n = 218,856 datapoints; right: Monkey V, n = 118,629 datapoints; each datapoint corresponds to a single neuron-condition pairing within an experiment). In both monkeys, the mean-matched Fano factor undergoes a transient reduction locked to the onset of the stimulus and the onset of the saccade. The reduction in variability around the time of saccade onset coincides with a contraction of the eigenvalues of the residual dynamics (Fig. 4b,e), suggesting that more quickly decaying dynamics may underlie variability quenching at that time. A contraction of eigenvalues, however, does not appear necessary to explain variability quenching, as an analogous contraction is not observed at the time of stimulus onset, despite the consistent reduction in variability at stimulus onset. **c,** Overall fraction of variance explained by the dynamics subspace. We quantified what fraction of the variance of the condition-averaged trajectories in the high-dimensional neural space (state space defined by the individual units) is contained in the dynamics subspace ($\boldsymbol{U}_{dyn}$, estimated in Step 2 of the analysis pipeline; Extended Data Fig. 2). Data from all 164 experiments in monkey T. On average in monkey T, the 8-dimensional dynamics subspace explains 68% of the variance in the average neural trajectories in monkey T (dashed vertical line, n = 164 experiments).*



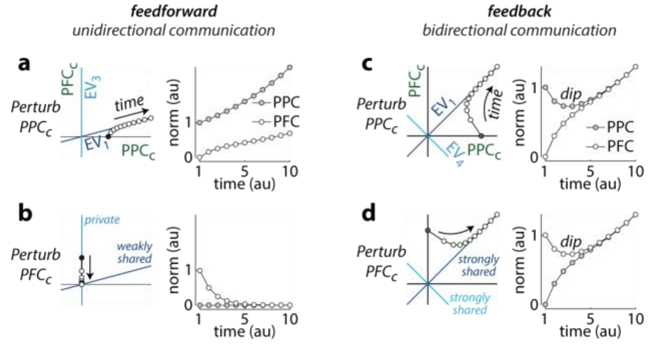
***Extended Data Fig. 8: Cross-validation of hyper-parameters used for estimating residual dynamics***

***a-c**, Representative results of the cross-validation procedure used to determine the various hyper-parameters of the analysis pipeline (Extended Data Fig. 2; see Methods) for neural data from a single task configuration in monkey T (config-3, see Extended Data Fig. 6a). **a**, Cross-validated hankel matrix reconstruction error ($E_{hankel}$; circle: mean over n = 20 repeats of hold-out cross validation; error bars: 1 s.e.m) plotted as a function of the rank of the hankel matrix (r, step 2 in Extended Data Fig. 2; see Methods) for residuals from the two epochs (left: decision; right: movement) and two choices (blue: choice 1; red: choice 2). The reconstruction error for each of the 20 repeats was computed by assigning a random 50% of the trials as a "training" set and the rest as a "test" set. **b**, 5-fold cross-validated mean squared error ($E_{fs}$; circles: mean over n=5 folds; error bars: 1 s.e.m) of the denoised residual predictions obtained from the first stage of the two-stage least squares regression (2SLS; step 3 in Extended Data Fig. 2), plotted as a function of the hyper-parameters: d (dimensionality of dynamics subspace); and l (number of past lags). For each cross-validation fold, a single mean squared error measure was computed by pooling the denoised predictions across time points in both epochs (left: choice 1; right: choice 2). **c**, Cross-validated mean squared error (circle: mean across n = 5 'repeats' of the average mean squared error across 5-folds; error bars: 2 std across repeats) of the residual predictions obtained from the second stage of the 2SLS regression (step 4 in Extended Data Fig. 2), plotted as a function of the smoothness hyper-parameter α for different epochs (left: decision; right: movement) and choice (choice 1 and 2). Both the train (orange) and test (gray) error are shown. **d**, Summary showing the optimal value for the dimensionality d and lag l (step 3 in Extended Data Fig. 2) for all task configurations and monkeys (symbols as in Extended Data Fig. 6a). A dimensionality of 8 and a lag of 3 was deemed optimal for both monkeys and task configurations (used in Fig 4). **e**, Summary showing the optimal smoothness hyper-parameter α (step 4 in Extended Data Fig. 2) for all task configurations and monkeys. Final values of α were chosen to be the same across monkeys in Fig. 4 (decision epoch: α = 200; movement epoch: α = 50) despite a small degree of variability across the two monkeys. Same conventions as in **d**.*



***Extended Data Fig. 9: Assessing statistical bias of eigenvalue estimates***

*We estimated the residual dynamics for different choices of bin size, to identify the smallest bin size resulting in unbiased estimates. In the discrete time formulation of a linear dynamical system, like the one we use here, re-binning of the responses trivially results in a scaling of the estimated eigenvalues of the residual dynamics. To compensate for this rescaling, here we "mapped" the estimated eigenvalues onto a common, reference bin size (see Methods). In the absence of statistical biases, the resulting "re-binned eigenvalue" would be independent of bin size.* ***a****, Re-binned eigenvalues for simulations of a time-invariant, latent-variable (3 latent dimensions), LDS model (reference bin size = 40ms) as a function of bin-size (dashed line: ground truth). Different gray lines correspond to models with different levels of latent noise (legend). When latent noise is large, estimates of the residual dynamics are biased for small bin sizes, but become unbiased when bin size is sufficiently large (light gray). When latent noise is too small, estimates are biased for any choice of bin size (black).* ***b****, Estimated, re-binned eigenvalues (reference bin size = 15ms) as a function of bin size for all configurations in monkey T. Columns correspond to the 8 distinct eigenmodes of the estimated 8-dimensional residual dynamics (left to right, largest to smallest EV), rows correspond to task configurations (top to bottom, config-1 to 4; see Extended Data Fig. 6a). Here the re-binned eigenvalues were computed separately for each choice (red vs blue) and averaged in small temporal windows specific to each epoch: 0.2-0.4s relative to stimulus onset (solid lines) and -0.15 to 0.25s relative to saccade onset (dashed lines). All main analyses of recorded neural responses are based on a bin size of 45ms, for which eigenvalue estimates have converged to an asymptote, suggesting that our estimates are not biased. Note that the re-binned eigenvalues for a bin size of 45ms are larger than the corresponding eigenvalues reported in other figures (e.g. Fig. 4b), because the former were mapped onto a reference bin size of 15ms.*



***Extended Data Fig. 10: Unidirectional and bidirectional communication between areas.***

*A population level mechanism explaining unidirectional and bidirectional communication between areas, incorporating key properties of the global residual dynamics in the feedforward (**a**, **b**) and feedback networks (**c**, **d**) in Fig. 7. We simulated time-independent, two-dimensional, linear dynamics, whereby the two cardinal dimensions (left panels in **a**-**d**) represent the choice modes in PPC and PFC (Fig. 6a,d). The time modes in each area are ignored here. We simulated a local perturbation (right panels in **a**-**d**) either in PPC (**a**, **c**) or PFC (**b**, **d**) by initializing activity along the corresponding choice mode (black circles, left panels) and then letting activity evolve (white points) based on the linear dynamics determined by the respective EVs (Fig 7a; see Supplementary Methods). **a**, Perturbation in PPC in the feedforward model. Left: evolution of activity in the two-dimensional, global state-space spanned by PPC and PFC. Right: time-course of the norm of the population activity. The PPC perturbation causes expanding activity in PPC that propagates to PFC. **b**, Perturbation of PFC in the feedforward model in Fig 6a. The PFC perturbation decays in PFC and does not propagate to PPC. This unidirectional communication results from non-normal dynamics, as EV$_1$ is shared, while EV$_3$ is private to PFC (EV$_1$ not orthogonal to EV$_3$). **c**, Perturbation of PPC in the feedback model. The PPC perturbation causes a dip in PPC and expanding activity in PFC. **d**, Perturbation of PFC in the feedback model in Fig 6d. The PFC perturbation causes a dip in PFC and expanding activity in PPC. In the feedback model, perturbations in one area thus propagate to the other area. This bidirectional communication arises because both EV$_1$ and EV$_4$ are shared equally between PPC and PFC. Somewhat counter-intuitively, the existence of bidirectional communication in these models can be inferred when considering activity in the perturbed area alone. Activity in the perturbed area initially decays, and expands only later; activity in the unperturbed area does not show this dip. The dip occurs because any local perturbation is only partially aligned with the shared, unstable direction (EV$_1$). Initially, activity in the perturbed area then mostly reflects the rapidly decaying component of activity along the second, global eigenvector (EV$_4$).*

# References

1. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).
2. Yuste, R. From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**, 487–497 (2015).
3. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
4. Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *J. Neurophysiol.* **102**, 614–635 (2009).
5. Linderman, S. W. *et al.* Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. *Proc. 20th Int. Conf. Artif. Intell. Stat.* **54**, 914–922 (2017).
6. Zhao, Y. & Park, I. M. Variational Latent Gaussian Process for Recovering Single-Trial Dynamics from Population Spike Trains. *Neural Comput.* **29**, 1293–1316 (2017).
7. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
8. Duncker, L., Bohner, G., Boussard, J. & Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. in *Proceedings of the 36th International Conference on International Conference on Machine Learning* vol. 97 1726–1734 (PMLR, 2019).
9. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat. Neurosci.* **22**, 1512–1520 (2019).
10. Mazor, O. & Laurent, G. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* **48**, 661–673 (2005).
11. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation Through Neural Population Dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
12. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

13. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
14. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian Computation through Cortical Latent Dynamics. *Neuron* **103**, 934-947.e5 (2019).
15. Barak, O., Sussillo, D., Romo, R., Tsodyks, M. & Abbott, L. F. From fixed points to chaos: Three models of delayed discrimination. *Prog. Neurobiol.* **103**, 214–222 (2013).
16. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99**, 609–623 (2018).
17. Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810-824.e9 (2019).
18. Sauerbrei, B. A. *et al.* Cortical pattern generation during dexterous movement is input-driven. *Nature* **577**, 386–391 (2020).
19. Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–96 (1998).
20. Churchland, A. K. *et al.* Variance as a Signature of Neural Computations during Decision Making. *Neuron* **69**, 818–831 (2011).
21. Churchland, M. M. *et al.* Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).
22. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–819 (2011).
23. Goris, R. L. T., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014).
24. Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E. & Doiron, B. The spatial structure of correlated neuronal variability. *Nat. Neurosci.* **20**, 107–114 (2017).
25. Ebrahimi, S. *et al.* Emergent reliability in sensory cortical coding and inter-area communication. *Nature* **605**, 713–721 (2022).
26. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).

27. Chettih, S. N. & Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in V1. *Nature* **567**, 334–340 (2019).

28. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).

29. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

30. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

31. Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nat. Neurosci.* **23**, 260–270 (2020).

32. Buesing, L., Sahani, M. & Macke, J. H. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. in *Advances in Neural Information Processing Systems* vol. 25 1682–1690 (2012).

33. Sani, O. G., Abbaspourazad, H., Wong, Y. T., Pesaran, B. & Shanechi, M. M. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nat. Neurosci.* **24**, 140–149 (2021).

34. Angrist, J. D. & Krueger, A. B. Instrumental variables and the search for identification: From supply and demand to natural experiments. *J. Econ. Perspect.* **15**, 69–85 (2001).

35. Wang, X.-J. Decision Making in Recurrent Neuronal Circuits. *Neuron* **60**, 215–234 (2008).

36. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

37. Laje, R. & Buonomano, D. V. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. Neurosci.* **16**, 925–933 (2013).

38. Murray, J. D., Jaramillo, J. & Wang, X.-J. Working Memory and Decision-Making in a Frontoparietal Circuit Model. *J. Neurosci.* **37**, 12167–12186 (2017).

39. Das, A. & Fiete, I. R. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nat. Neurosci.* **23**, 1286–1296 (2020).

40. Wilting, J. & Priesemann, V. Inferring collective dynamical states from widely unobserved systems. *Nat. Commun.* **9**, 2325 (2018).

41. Chaudhuri, R., Knoblauch, K., Gariel, M. A., Kennedy, H. & Wang, X.-J. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron* **88**, 419–431 (2015).

42. Gold, J. I. & Shadlen, M. N. The Neural Basis of Decision Making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).

43. Machens, C. K., Romo, R. & Brody, C. D. Functional, But Not Anatomical, Separation of 'What' and 'When' in Prefrontal Cortex. *J. Neurosci.* **30**, 350–360 (2010).

44. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* **102**, 249-259.e4 (2019).

45. Buesing, L., Macke, J. H. & Sahani, M. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. *Adv. Neural Inf. Process. Syst. NIPS* 1–9 (2012) doi:10.3109/0954898X.2012.677095.

46. Murphy, B. K. & Miller, K. D. Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. *Neuron* **61**, 635–648 (2009).

47. Goldman, M. S. Memory without Feedback in a Neural Network. *Neuron* **61**, 621–634 (2009).

48. Ganguli, S., Huh, D. & Sompolinsky, H. Memory traces in dynamical systems. *Proc. Natl. Acad. Sci.* **105**, 18970–18975 (2008).

49. Gao, P. *et al.* A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* (2017) doi:10.1101/214262.

50. Kiani, R. *et al.* Natural grouping of neural responses reveals spatially segregated clusters in prearcuate cortex. *Neuron* **85**, 1359–1373 (2015).

51. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).

52. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: Permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).

53. Javadzadeh, M. & Hofer, S. B. Dynamic causal communication channels between neocortical areas. *Neuron* **110**, 2470-2483.e7 (2022).

54. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).

55. Hart, E. & Huk, A. C. Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network. *eLife* **9**, e52460 (2020).

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

57. Kiani, R., Hanks, T. D. & Shadlen, M. N. Bounded Integration in Parietal Cortex Underlies Decisions Even When Viewing Duration Is Dictated by the Environment. *J. Neurosci.* **28**, 3017–3029 (2008).

58. Aoi, M. C., Mante, V. & Pillow, J. W. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nat. Neurosci.* **23**, 1410–1420 (2020).

59. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).

60. Hennequin, G., Vogels, T. P. & Gerstner, W. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron* **82**, 1394–1406 (2014).

61. Durstewitz, D. & Seamans, J. K. Beyond bistability: Biophysics and temporal dynamics of working memory. *Neuroscience* **139**, 119–133 (2006).

62. Deco, G. & Jirsa, V. K. Ongoing Cortical Activity at Rest: Criticality, Multistability, and Ghost Attractors. *J. Neurosci.* **32**, 3366–3375 (2012).

63. Dahmen, D., Grün, S., Diesmann, M. & Helias, M. Second type of criticality in the brain uncovers rich multiple-neuron dynamics. *Proc. Natl. Acad. Sci.* **116**, 13051–13060 (2019).

64. Purcell, B. A., Heitz, R. P., Cohen, J. Y. & Schall, J. D. Response variability of frontal eye field neurons modulates with sensory input and saccade preparation but not visual search salience. *J. Neurophysiol.* **108**, 2737–2750 (2012).

65. Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M. & Miller, K. D. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron* **98**, 846-860.e5 (2018).

66. Litwin-Kumar, A. & Doiron, B. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.* **15**, 1498–1505 (2012).

# Methods

## Experimental Procedures

All surgical, behavioral, and animal-care procedures complied with National Institutes of Health guidelines and were approved by the Stanford University Institutional Animal Care and Use Committee.

## Behavioral Task

Two adult male rhesus macaque monkeys (monkey T: 14kg, monkey V: 11kg) discriminated the direction of motion of a random dot motion kinetogram and reported their choice by saccades to one of two choice targets[67] (Fig. 3a). Visual stimuli were presented on a cathode ray tube monitor (viewing distance = 57 cm, frame rate = 120Hz) controlled by a VSG graphics card (Cambridge Graphics, UK). Each trial began with the appearance of a small spot requiring fixation for a duration of 500ms (± 1.5° visual angle, fixation window). Eye position was measured with a scleral search coil (CNC Engineering, Seattle, WA). The fixation period was followed by the appearance of two saccade targets (eccentricity 6-18°, angular locations varied across recording sessions) . After a 400ms delay, the random-dot stimulus was presented centered on the fixation point (circular aperture diameter: 7°/6°, monkey T/V) for a fixed duration of 800ms (decision epoch). The percentage of dots moving coherently in the same direction (motion coherency) controlled the task difficulty, and was chosen randomly on each trial from a fixed set of values. The decision epoch was followed by a delay period (no random dots; only fixation point and saccade targets visible) of variable duration (300-1100ms, mean = 700ms). Saccades were initiated following a 'Go' cue (disappearance of fixation point at end of the delay), followed by a 'hold' period (500-1200ms, mean 900ms) requiring fixation on the target (± 2-4° fixation window, depending on eccentricity). At the end of the hold period, both targets disappeared and a liquid reward was dispensed for correct trials (0% motion coherence trials rewarded at random).

## Neural Recordings

Single and multi-unit neural activity was recorded in the left cerebral hemisphere of both monkeys using surgically implanted[68], multi-channel electrode arrays (Blackrock Microsystems, Salt Lake City, UT) (96 electrodes; length = 1.5 mm; spacing = 0.4 mm) in the pre-arcuate gyrus (Brodmann's area 8Ar) between the posterior end of the principal sulcus and the anterior bank of the arcuate sulcus (Fig. 3b). Array signals were amplified with respect to a common subdural ground, filtered and digitized prior to spike-sorting. For each electrode, spikes from the entire duration of a recording session were sorted and clustered offline (Plexon Inc., Dallas, Texas) based on a principal component analysis of voltage waveforms. Candidate action-potential classifications for each electrode were subject to additional quality controls, including considerations of waveform shape, waveform reproducibility, inter-spike interval statistics, and the overall firing rate. Spike-sorting yielded approximately 100-200 single and multi-unit clusters distributed across the array in each recording session. The term "units" collectively refers to both isolated single-units and putative multi-units.

## Data pre-processing

We consider neural data in two non-overlapping time epochs of the trial: -200 to +1000ms relative to random-dots onset, or -700 to +500ms relative to movement (saccade) onset. In each recording session, we removed 'silent' units that had an average firing rate (computed across all trials and time-points) of $<$1Hz. For unknown reasons, in most sessions the neural data exhibited abrupt synchronous changes in the overall firing rate of many units locked to specific trial indices in the session. We automatically identified these putative 'change-points' and used them to split each recording session into shorter "experiments" (total number of experiments = 164/80 from 81/76 separate recording sessions, resulting in 58,187/34,451 trials in monkey T/V), within which the overall firing rate was stationary. Experiments with fewer than 200 trials were excluded from further analysis. We also removed units that exhibited strong discontinuities in their temporally averaged firing rate across trials, within every experiment. Square-root transformed binned spike-counts[4] were computed for each unit in non-overlapping time bins (45ms long, see Extended Data Fig. 9).

Data from each experiment was assigned to one of four different "task configurations" based on the coarse angular positions of the two choice targets (Extended Data Fig. 6a). Each trial was categorized either as a choice 1 or choice 2 trial depending on the selected target. In 3 out of 4 task configurations, choice-1 corresponds to saccades to the contra-lateral visual hemifield (blue targets, Extended Data Fig. 6a). For each experiment we computed the percentage of responses to the choice 1 target as a function of signed motion coherence and, fitted a logistic sigmoidal curve to all the resulting data points that came from the same task configuration (Extended Data Fig. 6b).

## Overview of the analysis

Assuming simple inputs (see main text, Fig. 1b), an analysis of response residuals can reveal the properties of recurrent dynamics $\mathbf{F}(.)$, even when input $\mathbf{u}_t$ is unknown (Equation 1). Henceforth, for simple inputs we assume (without loss of generality) that the input equals the input drive ($\mathbf{u}_t \equiv \bar{\mathbf{u}}_t$, $\bar{\mathbf{u}}_t$ defined as input drive in main text; Fig. 1b, Supplementary Methods), whereby the uncorrelated input latent noise is implicitly included within the latent noise ($\boldsymbol{\epsilon}_t$) in Equation 1.

The instantaneous change in the single trial state on trial *k* at time *t* is given by:

$$\dot{\mathbf{z}}_t^k = \mathbf{F}\left(\mathbf{z}_t^k\right) + \mathbf{u}_t + \boldsymbol{\epsilon}_t \tag{2}$$

Likewise, we assume that the instantaneous change in the average state across trials (denoted by $\langle . \rangle$) can be written as:

$$\langle \dot{\mathbf{z}}_t \rangle = \mathbf{F}(\langle \mathbf{z}_t \rangle) + \mathbf{u}_t \tag{3}$$

The equality in the above equation follows from Equation 2 (as $\boldsymbol{\epsilon}_t$ is zero-mean), in particular if $\mathbf{F}$ is locally linear. The average state and the $k^{\text{th}}$ single trial state are approximated by the following discretized updates

$$\langle \mathbf{z}_{t+1} \rangle \quad = \langle \mathbf{z}_t \rangle + \Delta t \cdot (\mathbf{F}(\langle \mathbf{z}_t \rangle) + \mathbf{u}_t) \tag{4}$$

$$\mathbf{z}_{t+1}^k \ = \ \mathbf{z}_t^k + \Delta t \cdot \left(\mathbf{F}\left(\mathbf{z}_t^k\right) + \mathbf{u}_t + \ \boldsymbol{\epsilon}_t\right) \tag{5}$$

The dynamics of the residual vector $\tilde{\mathbf{z}}$ on the $k^{\text{th}}$ trial is obtained as

$$\underbrace{\mathbf{z}_{t+1}^k - \langle \mathbf{z}_{t+1} \rangle}_{=\tilde{\mathbf{z}}_{t+1}^k} \quad = \underbrace{\mathbf{z}_t^k - \langle \mathbf{z}_t \rangle}_{=\tilde{\mathbf{z}}_t^k} + \Delta t \cdot \left(\mathbf{F}\left(\mathbf{z}_t^k\right) - \mathbf{F}(\langle \mathbf{z}_t \rangle) + \ \boldsymbol{\epsilon}_t\right) \tag{6}$$

Therefore, the temporal evolution of the residuals is itself governed by a differential equation, expressed in terms of the single trial dynamics as

$$\underbrace{\dot{\tilde{\mathbf{z}}}_t}_{\text{residual flow}} \quad = \ \underbrace{\left(\mathbf{F}\left(\mathbf{z}_t^k\right) + \mathbf{u}_t\right)}_{\text{"effective" flow at } \mathbf{z}_t^k} \ - \ \underbrace{\left(\mathbf{F}(\langle \mathbf{z}_t \rangle) + \mathbf{u}_t\right)}_{\text{"effective" flow at } \langle \mathbf{z_t} \rangle} \ + \ \boldsymbol{\epsilon}_t \tag{7}$$

Grouping and rearranging terms of Equation 6, we obtain.

$$\mathbf{F}\left(\mathbf{z}_t^k\right) - \mathbf{F}(\langle \mathbf{z}_t \rangle) + \ \boldsymbol{\epsilon}_t \quad = \mathbf{F}\left(\langle \mathbf{z}_t \rangle + \tilde{\mathbf{z}}_t^k\right) - \mathbf{F}(\langle \mathbf{z}_t \rangle) + \ \boldsymbol{\epsilon}_t \tag{8}$$

A first order Taylor expansion of the first term on the right-hand side of Equation 8 results in:

$$\mathbf{F}\left(\langle \mathbf{z}_t \rangle + \tilde{\mathbf{z}}_t^k\right) = \mathbf{F}(\langle \mathbf{z}_t \rangle) + \underbrace{\nabla \mathbf{F}|_{\langle \mathbf{z}_t \rangle}}_{=\mathbf{J}_t} \ . \ \tilde{\mathbf{z}}_t^k + \text{higher order terms} \tag{9}$$

Ignoring second and higher order terms, and re-expressing Equation 6 using Equations 8 and 9 yields a discrete-time, time-varying, linear dynamical system at the level of the residuals

$$\begin{aligned} \tilde{\mathbf{z}}_{t+1} \quad &= \tilde{\mathbf{z}}_t + \Delta t . (\mathbf{J}_t \tilde{\mathbf{z}}_t + \ \boldsymbol{\epsilon}_t) \\ &= \Big(\underbrace{\mathbf{I} + \Delta t . \mathbf{J}_t}_{=\mathbf{A}_t}\Big) \tilde{\mathbf{z}}_t + \ \Delta t . \boldsymbol{\epsilon}_t \end{aligned} \tag{10}$$

The time-varying "dynamics matrix" ($\mathbf{A}_t$) maps residuals from time *t* to *t+1*, and is directly related to the Jacobian ($\mathbf{J}_t$) of the underlying dynamical system, computed at each state along the average trajectory. Critically, $\mathbf{u}_t$ does not appear in Equation 10, meaning that, for simple inputs (Fig. 1b) and instantaneous noise $\boldsymbol{\epsilon}_t$, the residual dynamics $\mathbf{A}_t$ only reflects the recurrent dynamics. The corresponding analytical derivations for the complex input regime (Fig. 1b) are considered in Supplementary Math Note B.

Residuals obtained from neural population spike-counts are modeled using a latent-variable, autonomous, linear, time-varying dynamical system as described below

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}_t \mathbf{x}_t + \boldsymbol{\epsilon}_t \\ \tilde{\mathbf{z}}_t &= \mathbf{C} \mathbf{x}_t + \boldsymbol{\eta}_t \end{aligned} \tag{11}$$

where, $\mathbf{x}_t$ is a low-dimensional, *latent residual* state with dynamics determined by $\mathbf{A}_t$ (Equation 10-11), and is mapped linearly through an “observation matrix” ($\mathbf{C}$) resulting in observed residuals $\tilde{\mathbf{z}}_t$ . $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are “latent” and “observation” gaussian noise processes. The subspace spanned by the columns of $\mathbf{C}$, termed the “dynamics subspace” (later denoted by $\mathbf{U}_{dyn}$, Extended Data Fig. 2), contains the *dynamically relevant* portion of response variability (trial-by-trial variability along any dimension within it covaries with variability along the same or other dimensions at later times). The observations $\tilde{\mathbf{z}}_t$ could, in principle, directly correspond to the neural spike-count residuals (Extended Data Fig. 7). However, in most of our analyses, they correspond to a low dimensional projection of neural spike-count residuals, obtained by aligning neural data across multiple experiments (Extended Data Fig. 2).

Linear, time-varying latent dynamics (Equation 11) make exact probabilistic inference intractable, requiring approximate inference techniques[5,6,8]. We estimate parameters $\mathbf{C}$ and $\mathbf{A}_t$ using an alternative approach, combining subspace system identification (SSID) theory[33,45,69] (see Supplementary Math Note A) and instrumental variable (IV) regression[34,70]. For PFC responses, the amount of variance explained by the inferred dynamics appear to be limited primarily by the large contribution of (unpredictable) observation noise (see below, Qualitative estimates of goodness of fit).

## Neural data analysis pipeline

We developed a data analysis pipeline (Extended Data Fig. 2) to estimate the dynamics subspace ($\mathbf{U}_{dyn}$) and the residual dynamics ($\mathbf{A}_t$) in four steps: (i) aligning neural responses across different experiments (**session alignment**), (ii) using aligned residuals pooled across experiments to estimate the dynamics subspace (**dynamics subspace estimation**), (iii) using aligned residuals and the dynamics subspace to estimate the latent residual state $\mathbf{x}_t$ (**residual latent state estimation**) and, (iv) combining the outputs of the previous three steps to estimate the residual dynamics $\mathbf{A}_t$

(**time-varying dynamics estimation**). We used SSID in Step (ii) and two-stage least squares (2SLS) based on instrumental variables for Steps (iii) & (iv)

### *Session alignment*

We aligned condition-averaged neural activity from different experiments (Extended Data Fig. 2, Step 1) to improve the statistical power of our analyses, assuming that neural population activity in different experiments correspond to different high-dimensional readouts of a fixed set of low-dimensional activity patterns[31]. A full example of the results of the alignment procedure applied to neural data from a single task configuration in one monkey is shown in Extended Data Fig. 6c-g.

We constructed (separately for each task configuration) a block condition average matrix ($\overline{\mathbf{Y}}_{\text{joint}}$) by concatenating row-wise, the trial averaged, neural population binned spike-counts ($\overline{\mathbf{Y}}_i$) of each experiment:

$$\overline{\mathbf{Y}}_{\text{joint}} = \begin{pmatrix} \overline{\mathbf{Y}}_1 \\ \overline{\mathbf{Y}}_2 \\ \vdots \\ \overline{\mathbf{Y}}_P \end{pmatrix} = \mathbf{U}.\mathbf{S}.\mathbf{V}' \tag{12}$$

where, $\overline{\mathbf{Y}}_i$ is a $\mathrm{N}_i \times (\mathrm{T}_{\text{all}} \times \mathrm{C})$ data matrix (mean centered; $\boldsymbol{\mu}_i$= subtracted row means, $\mathrm{N}_i$ = number of units for experiment *i*; $\mathrm{T}_{\text{all}}$ = total number of time bins in the decision and movement epochs; C = total number of conditions; *P* = total number of experiments to be aligned) computed by averaging single-trial trajectories, sorted into two choice conditions (C = 2, choice 1 or choice 2).

The singular value decomposition (SVD) of $\overline{\mathbf{Y}}_{\text{joint}}$ resulted in a matrix of left singular vectors ($\mathbf{U}$ in Equation 12), block-structured, of size $\sum_{i=1}^{\mathrm{P}} \mathrm{N}_i \times (\mathrm{T}_{\text{all}} \times \mathrm{C})$, represented as:

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1^1 & \mathbf{u}_1^2 & \dots & \mathbf{u}_1^{\mathrm{T.C}} \\ \mathbf{u}_2^1 & \mathbf{u}_2^2 & \dots & \mathbf{u}_2^{\mathrm{T.C}} \\ \vdots & & \ddots & \vdots \\ \mathbf{u}_P^1 & \mathbf{u}_P^2 & \dots & \mathbf{u}_P^{\mathrm{T.C}} \end{bmatrix} \tag{13}$$

where, $\mathbf{u}_i^j$, is the left singular 'sub-vector' (size-$\mathrm{N}_i \times 1$) corresponding to mode *j* in experiment *i*. The aligned coordinate basis, defined as matrix $\mathbf{U}_{i,\mathrm{M}}^{\perp}$ (size $\mathrm{N}_i \times \mathrm{M}$) for experiment *i*, corresponds to the first M orthogonalized columns (QR decomposition) of the *i*th block row of $\mathbf{U}$.

The M-dimensional, aligned single trial response $\mathbf{z}_t^i(k)$, at time *t* on trial *k* in experiment *i* is obtained as:

$$\mathbf{z}_t^i(k) \quad = \mathbf{U}_{i,\mathrm{M}}^{\perp}{}'.\left(\mathbf{y}_t^i(k) - \boldsymbol{\mu}_i\right) \tag{14}$$

where, $\mathbf{y}_t^i(k)$ is the corresponding neural spike-count population vector (size $\mathrm{N}_i \times 1$). This procedure therefore resulted in *P* aligned, single trial data matrices $\mathbf{Z}^i$ (*i* = 1, 2, …, *P*), each of size $\mathrm{M} \times \mathrm{T}_{\mathrm{all}} \times \mathrm{K}_i$ , where $\mathrm{K}_i$ is the number of trials in experiment *i*.

We inspected the cumulative amount of variance explained in the condition-averaged data matrix $\overline{\mathbf{Y}}_i$ as a function of M, by progressively retaining a larger number of columns for constructing $\mathbf{U}_{i,\mathrm{M}}^{\perp}$ . The fraction of variance explained by M aligned modes in experiment *i* is given by:

$$\left(1 - \frac{\mathrm{var}\ \left(\overline{\mathbf{Y}}_i - \mathbf{U}_{i,\mathrm{M}}^{\perp}.\mathbf{U}_{i,\mathrm{M}}^{\perp}{}'.\overline{\mathbf{Y}}_i\right)}{\mathrm{var}\left(\overline{\boldsymbol{Y}}_i\right)}\right) \tag{15}$$

For all subsequent analyses, we chose M = 20 (Extended Data Fig. 6c). We visualized each of the twenty aligned activity modes, obtained for experiment *i* by projecting $\overline{\mathbf{Y}}_i$ into $\mathbf{U}_{i,20}^{\perp}$, either individually for each experiment (Extended Data Fig. 6e), or by averaging across experiments (Extended Data Fig. 6d).

To evaluate the efficacy of alignment, we computed a correlation coefficient Corr ($< \mathbf{z}_{(a)}^i >$ , $<\mathbf{z}_{(b)}^j >$) for any given pair of aligned modes (indexed by a, *b*, where *a,b* ∈ {1,2,..,20}) across all possible pairs of experiments (indexed by *i* and *j*), where $< \mathbf{z}_{(a)}^i >$ (size $1 \times (\mathrm{T}_{\mathrm{all}} \times \mathrm{C})$) is the trial-averaged, activity time course (for both choices) of the *a*th aligned mode in experiment *i*. We then computed the median correlation coefficient across all pairs of dissimilar experiments (*i≠j)* for each pair of modes and visualized the resulting correlation matrix. The median correlation coefficient matrix (Extended Data Fig. 6g) displayed large values along the diagonal and almost zero values along the off-diagonals indicating that the aligned time-courses were much more similar across sessions than across modes.

### *Dynamics Subspace Estimation*

We estimated the dynamics subspace (Extended Data Fig. 2, Step 2) using residuals computed in the 20-dimensional space of aligned activity patterns. The dynamics subspace was estimated using SSID which is based on the idea of finding “temporally-predictive” directions in state space[69] (see Supplementary Math Note A). We adapted existing SSID methods for linear time-invariant systems, to make them suitable for linear time-varying dynamics (Equation 11).

To compute residuals, we *redefined* conditions separately for the two task epochs (decision/movement). For the decision epoch, we defined conditions based on choice *and* motion strength (2 choices × 4 to 8 coherencies ≈ 8 to 16 conditions; number of distinct motion coherencies varied across different experiments). For the movement epoch, we defined conditions based on choice and the length of the delay period preceding the 'go' cue, by sorting trials in each experiment into 5 different groups based on the length of the delay period (bin boundaries = [0 0.4 0.6 0.8 1.0 1.5]s). To ensure minimal overlap between the decision and movement epochs, we excluded all trials with delay lengths < 400ms. For the movement epoch, we obtained a total of 8 conditions in monkey T (2 choices × 4 delay length bins) and 6 conditions in monkey V (2 choices × 3 delay length bins; no trials in monkey V had delays > 1s) across all experiments. For each condition, we subtracted from the aligned single trial trajectories ($\mathbf{Z}^i$) the corresponding condition-averaged trajectory, which ultimately resulted in *P* aligned residual data matrices $\tilde{\mathbf{Z}}^i$ (*i* = 1, 2, …, *P*; each of size $20 \times \mathrm{T}_{\mathrm{all}} \times \mathrm{K}_{\mathrm{i}}$) for each experiment *i*. We then sorted trials in each $\tilde{\mathbf{Z}}^i$ based on choice (choice 1 or choice 2), and pooled them across the *P* experiments, resulting in two, choice-dependent, 'pooled' residual data matrices, $\tilde{\mathbf{Z}}_{\{\mathrm{choice}=1\}}$and $\tilde{\mathbf{Z}}_{\{\mathrm{choice}=2\}}$. All subsequent procedures were carried out *separately* on $\tilde{\mathbf{Z}}_{\{\mathrm{choice}=1\}}$ and $\tilde{\mathbf{Z}}_{\{\mathrm{choice}=2\}}$. For sake of convenience, below we drop the subscripts unless otherwise indicated.

Based on SSID theory, we constructed a sequence of *time-varying,* future-past hankel covariance matrices ($\mathbf{H}_t$) using temporally windowed chunks of $\tilde{\mathbf{Z}}$ centered at time *t* (Supplementary Math Note A, Equation S14). Specifically, we assigned trials in $\tilde{\mathbf{Z}}$ to two random halves (labelled "train" and "test"), and constructed two distinct, corresponding hankel matrices $\mathbf{H}_t^{train}$ and $\mathbf{H}_t^{test}$, at each time *t*. The order of the hankel matrix (given by *q* in Equation S14, Supplementary Math Note A), which determines the number of "future" and "past" lags of $\tilde{\mathbf{Z}}$ to use for constructing $\mathbf{H}_t$, is set to 5. Increasing *q* beyond 5 did not change the results of our analyses. We obtained the *r*-rank approximation of ${\mathbf{H}_t}^{train}$ (Extended Data Fig. 2, Step 2) by using a hard-thresholding of its singular values:

$$\mathbf{H}_{t,(r)}^{train} \quad = \mathbf{U}_{t,(r)}^{train} . \mathbf{S}_{t,(r)}^{train} . \mathbf{V}_{t,(r)}^{train\prime} \tag{16}$$

where, $\mathbf{U}_{t,(r)}^{train}$ and $\mathbf{V}_{t,(r)}^{train}$ are matrices whose columns are the first *r* left and right singular vectors of ${\mathbf{H}_t}^{train}$ respectively. Similarly, $\mathbf{S}_{t,(r)}^{train}$ is a diagonal matrix, with diagonal entries corresponding to the first *r* singular values. We then computed a temporally averaged, hankel matrix reconstruction error with respect to the full rank hankel matrix computed using the "test" trials:

$$\mathrm{E}_{\mathrm{hankel}} = \frac{1}{T - 2q + 1} \sum_{t=q+1}^{T-q+1} \| \mathbf{H}_t^{test} - \mathbf{H}_{t,(r)}^{train} \|_F^2 \tag{17}$$

where $\|.\|_F$ is the matrix frobenius norm, and T is the total number of time bins in $\tilde{\mathbf{Z}}$ for a specific task epoch (either decision or movement). We computed $\mathrm{E_{hankel}}$ using 20 different random splits of $\tilde{\mathbf{Z}}$ into "train" and "test" halves, for different values of the hankel rank (*r*). The average reconstruction error (over 20 repeats) was plotted as a function of *r (*Extended Data Fig. 8a) and the optimal rank ($r_{opt}$) was determined as the smallest value of *r* for which $\mathrm{E_{hankel}}$ was no larger than one standard error above the minimum $\mathrm{E_{hankel}}$ (1 standard error rule[71]). Thus, we obtained a single value of $r_{opt}$ for each task epoch and choice condition. While a single $r_{opt}$ (determined across all times in an epoch) may over/under estimate the optimal rank at a specific time *t*, we found that using an $r_{opt}$ deemed optimal at each time *t* also yielded similar results.

Next, we used the above estimate of $r_{opt}$ and the aligned residuals to define observability matrices, which were eventually used to estimate the dynamics subspace. For the subsequent steps of the pipeline, we used a 5-fold cross validation approach. Time-varying hankel matrices ($\mathbf{H}_t$) were computed using $\tilde{\mathbf{Z}}^{train}$ (composed of 4/5th of all trials in $\tilde{\mathbf{Z}}$) and subjected to an SVD. The resulting, first $r_{opt}$ left singular vectors and singular values were used to define a time-dependent observability matrix $\widehat{\boldsymbol{\mathcal{O}}}_t$ (Equation S18, Supplementary Math Note A):

$$\widehat{\boldsymbol{\mathcal{O}}}_t \quad = \mathbf{U}_{t,(r_{opt})} . (\mathbf{S}_{t,(r_{opt})})^{\frac{1}{2}} \tag{18}$$

where, $\widehat{\boldsymbol{\mathcal{O}}}_t$ is a block matrix of size $(\mathrm{M} \times q) \times r_{opt}$, *q* = 5 is the order of the hankel matrix and M= 20 is the dimensionality of the aligned space. As in SSID for time-invariant dynamical systems, the first block row of $\widehat{\boldsymbol{\mathcal{O}}}_t$ specifies the *momentary* dynamics subspace at time *t*, given by the first M rows of $\widehat{\boldsymbol{\mathcal{O}}}_t$:

$$\hat{\mathbf{C}}_t \quad = \widehat{\boldsymbol{\mathcal{O}}}_t(1\text{:M}, :) \tag{19}$$

To define a single *time-invariant dynamics subspace* as in our model (Equation 11), from the sequence $\hat{\mathbf{C}}_t$ we constructed a matrix $\widehat{\mathbf{C}_{\mathrm{all}}}$ by concatenating, column-wise, the momentary dynamics subspaces $\hat{\mathbf{C}}_t$ for all *t* across both task epochs (decision and movement), and both choices (choice 1 or choice 2*)*

$$\widehat{\mathbf{C}_{\mathrm{all}}} = \left(\hat{\mathbf{C}}^{de,1}_{q+1} \ldots\ldots \hat{\mathbf{C}}^{de,1}_{T_{de}-q+1} \quad \hat{\mathbf{C}}^{mo,1}_{q+1} \ldots\ldots \hat{\mathbf{C}}^{mo,1}_{T_{mo}-q+1} \quad \hat{\mathbf{C}}^{de,2}_{q+1} \ldots\ldots \hat{\mathbf{C}}^{de,2}_{T_{de}-q+1} \quad \hat{\mathbf{C}}^{mo,2}_{q+1} \ldots\ldots \hat{\mathbf{C}}^{mo,2}_{T_{mo}-q+1}\right) \tag{20}$$

where, $\hat{\mathbf{C}}^{de,j}_t$ and $\hat{\mathbf{C}}^{mo,j}_t$ are the momentary dynamics subspaces for choice *j* (*j* = 1 or 2), at time *t* in the decision (*de*) and movement (*mo*) epochs. $T_{de}$ and $T_{mo}$ are the total number of time bins in the decision and movement epochs. The left singular vectors of $\widehat{\mathbf{C}_{\mathrm{all}}}$ , by definition, span the *union* of the column spaces of all $\hat{\mathbf{C}}_t$ (across time, task epochs and choice conditions), and therefore specify a time-invariant dynamics subspace *shared* across time, task epochs and choices. We denote the left-singular vectors of $\widehat{\mathbf{C}_{\text{all}}}$ by $\mathbf{U}_{dyn}$, which is an orthonormal matrix of size $\mathrm{M} \times \mathrm{M}$. The $\mathrm{M}$ columns of $\mathbf{U}_{dyn}$ are ordered in terms of their relative importance in capturing temporally correlated variability in the residuals resulting from the underlying dynamics (Fig. 4a, Extended Data Fig. 6f). In practice, only an ordered subset of the columns of $\mathbf{U}_{dyn}$ are sufficient to capture residual dynamics across choices and task epochs (using more columns than necessary leads to over-fitting). Accordingly, the number of columns of $\mathbf{U}_{dyn}$ that are retained corresponds to a hyper-parameter (denoted by *d*) that determines the dimensionality of the residual dynamics ($\mathbf{A}_t$ , Equation 11). We determine the optimal value of *d* (denoted as $d_{opt}$) using cross-validation in the next step of the pipeline (Extended Data Fig. 8). The first $d_{opt}$ columns of $\mathbf{U}_{dyn}$(Fig. 4a, Extended Data Fig. 6f) therefore correspond to the estimate of the observation matrix of our model (**C**, Equation 11).

### *Overview of two-stage least squares (2SLS)*

Next, we estimate the latent residual state ($\mathbf{x}_{\mathrm{t}}$ , Equation 11; Extended Data Fig. 2, Step 3) and the time-varying residual dynamics ($\mathbf{A}_t$ , Equation 11; Extended Data Fig. 2, Step 4) using a two-stage least squares (2SLS) approach based on instrumental variable regression.

First, we obtained a *d*-dimensional, *noisy* estimate of the latent residual state at time *t*, for each trial *k* in the training fold, by projecting the corresponding *observed* residual ($\tilde{\mathbf{z}}_t^{train}(k)$) into a *d*-dimensional dynamics subspace:

$$\tilde{\mathbf{x}}_t^{train}(k) = (\mathbf{U}_{dyn}^d)'\tilde{\mathbf{z}}_t^{train}(k) \tag{21}$$

where, $\mathbf{U}_{dyn}^d$ are the first *d* columns of $\mathbf{U}_{dyn}$ (estimated in Step 2 using only “train” trials, $\tilde{\mathbf{Z}}^{train}$). Such a projection does not entirely eliminate the observation noise present in $\tilde{\mathbf{z}}_t^{train}$; specifically, observation noise lying within the column space of $\mathbf{U}_{dyn}^d$ corrupts $\tilde{\mathbf{x}}_t$. Therefore, if one were to directly estimate residual dynamics ($\mathbf{A}_t$) using ordinary least squares (OLS) by regressing $\tilde{\mathbf{x}}_t$ against $\tilde{\mathbf{x}}_{t+1}$ (as suggested by Equation 11), the resulting estimates would be biased and inconsistent (Extended Data Fig. 3d). This is commonly referred to as the “error-in-variables” problem[72], in which components of observation noise corrupting $\tilde{\mathbf{x}}_t$ act as a confounding variable, resulting in an *attenuation bias* in OLS estimates of $\mathbf{A}_t$ (Extended Data Fig. 3d). Such biases would complicate the interpretation of the eigen/singular value spectrum of $\mathbf{A}_t$, which are crucial for drawing conclusions about underlying computations.

Therefore, we instead use an instrumental variable regression approach, commonly used to help mitigate the deleterious effects of confounding variables for causal inference[34], which relies on two separate least-squares regressions performed in two stages (2SLS). Two key assumptions underly the validity of this approach: (i) dynamics is considered markovian, and (ii) observation noise is considered temporally uncorrelated. In the first stage, we regress the noisy, latent residual state at time *t* ($\tilde{\mathbf{x}}_t$) against its past $l$ lags, $[\tilde{\mathbf{x}}_{t-1},\ \tilde{\mathbf{x}}_{t-2}\ \ldots \tilde{\mathbf{x}}_{t-l}]$. These lagged variables known as "instruments" (their validity subject to the above two assumptions), are therefore used to construct a *de-noised* prediction of the latent residual state at time *t* (Extended Data Fig. 2, Step 3). In the second stage, the noisy, latent residual state at time *t+1* ($\tilde{\mathbf{x}}_{t+1}$) is regressed against this *de-noised* prediction to obtain estimates of $\mathbf{A}_t$ that are unbiased and consistent (Extended Data Fig. 2, Step 4). 2SLS estimates of $\mathbf{A}_t$ can be potentially biased when instruments are 'weak'[73] (i.e when past lags have low predictive power in the first stage regression), underscoring the need to choose optimal values for the hyper-parameters *d* and $l$ (Extended Data Fig. 8b).

### *Residual latent state estimation*

The first stage of 2SLS involved estimating, at each time *t* (*separately* in each task epoch, for trials from the two choice conditions), the regression coefficients $\boldsymbol{\beta}_t^{l-}$ (using least squares) as follows:

$$\hat{\boldsymbol{\beta}}_t^{l-} = (\tilde{\mathbf{X}}_t^{train}\tilde{\mathbf{X}}_{t,l-}^{train\prime})(\tilde{\mathbf{X}}_{t,l-}^{train}\tilde{\mathbf{X}}_{t,l-}^{train\prime})^{-1} \tag{22}$$

where, $\tilde{\mathbf{X}}_t^{train}$ is a matrix of size $\mathrm{d} \times \mathrm{K}_{\mathrm{train}}$, whose columns correspond to the noisy latent residual state ($\tilde{\mathbf{x}}_t^{train}$ , Equation 21) for individual trials in the "training" set. Similarly, $\tilde{\mathbf{X}}_{t,l-}^{train}$ is a matrix of size $(\mathrm{d}\ \times\ \mathrm{l}) \times \mathrm{K}_{\mathrm{train}}$ , where each column corresponds to the past $l$ lags (stacked vertically) relative to $\tilde{\mathbf{x}}_t^{train}$, for the corresponding trial. Therefore, the *k*th column of $\tilde{\mathbf{X}}_{t,l-}^{train}$ (corresponding to trial index *k*) is a vector of size $(\mathrm{d}\ \times\ \mathrm{l}) \times 1$ specified as:

$$\tilde{\mathbf{x}}_{t,l-}^{train}(k) = \begin{bmatrix} \tilde{\mathbf{x}}_{t-1}^{train}(k) \\ \tilde{\mathbf{x}}_{t-2}^{train}(k) \\ \vdots \\ \tilde{\mathbf{x}}_{t-l}^{train}(k) \end{bmatrix} \tag{23}$$

We then predicted observed residuals in the test set ($\tilde{\mathbf{Z}}^{test}$, remaining 1/5th of the data) using estimates of $\mathbf{U}_{dyn}^{d}$ and $\hat{\boldsymbol{\beta}}_t^{l-}$ (both estimated using $\tilde{\mathbf{Z}}^{train}$) by first obtaining a noisy latent residual state at each time *t* for each trial in the test set (denoted by $\tilde{\mathbf{x}}_{\mathrm{t}}^{test}(k)$, analogous to Equation 21). The denoised prediction of the corresponding latent residual state is obtained as.

$$\hat{\tilde{\mathbf{x}}}_t^{test}(k) = \hat{\boldsymbol{\beta}}_t^{l-}\tilde{\mathbf{x}}_{t,l-}^{test}(k) \tag{24}$$

The corresponding prediction of the observed residual is then obtained by projecting $\hat{\tilde{\mathbf{x}}}_t^{test}(k)$ (Equation 24) back into the 20-dimensional aligned space, through the columns of $\mathbf{U}_{dyn}^{d}$:

$$\hat{\tilde{\mathbf{z}}}_t^{test}(k) = \mathbf{U}_{dyn}^{d}\hat{\tilde{\mathbf{x}}}_t^{test}(k) \tag{25}$$

These predictions were then used to compute a single mean squared error value for both task epochs (i.e. summation index *t* below spans both epochs) as follows:

$$E_{fs} = \frac{1}{T_{pred}.K_{test}} \sum_{k} \sum_{t} \| \hat{\tilde{\mathbf{z}}}_t^{test}(k) - \tilde{\mathbf{z}}_t^{test}(k) \|_2^2 \tag{26}$$

where, $T_{pred}$ corresponds to the total number of time-bins across both epochs (including only those time indices *t* which are greater than the maximum lag used for grid-search cross-validation), and $K_{test}$ is the total number of trials in $\tilde{\mathbf{Z}}^{test}$.

Different values of hyper-parameters *d* (dimensionality) and $l$ (number of past lags) were sampled on a two-dimensional grid. The resulting values of $E_{fs}$ (averaged across folds) for different settings of *d* and $l$ revealed a tendency to over-fit for large values (Extended Data Fig.8b). The optimal values of *d* and $l$ (denoted henceforth as $d_{opt}$and $l_{opt}$) were determined as the combination that resulted in the smallest number of parameters for $\hat{\boldsymbol{\beta}}_t^{l-}$(Equation 22), with an average $E_{fs}$ value no larger than one standard error above the minimum average $E_{fs}$ (1 standard error rule[71]).

### *Time-varying dynamics estimation*

For the second stage of 2SLS, first, we used optimal values ($d_{opt}$ and $l_{opt}$) of hyper-parameters *d* and $l$ (determined in the previous step), to recompute the optimal dynamics subspace ($\mathbf{U}_{dyn}^{d_{opt}}$; using all trials from both choices within a task configuration) and the optimal, denoised predictions of the latent residual states (Equation 24; using all trials of a specific choice and task configuration). To obtain residual dynamics ($\mathbf{A}_t$), we then solved (in closed form) the following penalized least squares objective[74]:

$$\mathcal{L} = \sum_{t} \| \tilde{\mathbf{X}}_{t+1} - \mathbf{A}_t \hat{\tilde{\mathbf{X}}}_t \|_F^2 + \alpha \| \mathbf{A}_{t+1} - \mathbf{A}_t \|_F^2 \tag{27}$$

where, $\tilde{\mathbf{X}}_{t+1}$ is a matrix whose columns correspond to the noisy ($d_{opt}$-dimensional), residual latent states at time *t+1*, for individual trials (obtained analogously as in Equation 21); $\hat{\tilde{\mathbf{X}}}_t$ is a matrix whose columns correspond to denoised predictions of the latent residual states at time *t* for corresponding trials ($k^{\text{th}}$ column corresponds to $\hat{\tilde{\mathbf{x}}}_t(k)$, analogous to Equation 24). The above objective is optimized separately for each task epoch (therefore, *t* in Equation 27 indexes only time bins within an epoch*)*. Critically, $\alpha$ is a regularization parameter (Extended Data Fig. 2, Step 4) that imparts smoothness (larger values implies more smoothness) to the sequence of dynamics matrices ($\mathbf{A}_t$) across time, and is tuned in a separate 5-fold cross-validation step. As expected, very small /large values of $\alpha$ exhibit over/under-fitting (Extended Data Fig. 8c).

### *Choices of hyperparameters*

We found that values of $d_{opt}$ = 8 and $l_{opt}$ = 3 were optimal for all 8 datasets (one dataset consists of trials for a specific choice and task configuration; Extended Data Fig. 8d) in each monkey. Despite small amounts of variability across different datasets, we used these fixed values for consistency, and to facilitate easier comparison of residual dynamics across different datasets. The 8-dimensional dynamics subspace $\mathbf{U}_{dyn}^{d_{opt}}$ computed using *only* residuals explained 68% (monkey T, median across 164 experiments; Extended Data Fig.7c) and 55% (monkey V, median across 80 experiments) of the variance in the trial-averaged trajectories in the high-dimensional neural space ($\overline{\mathbf{Y}}_i$, computed prior to alignment of sessions), as compared to 87% (monkey T) and 73% (monkey V) explained by the 20-dimensional aligned subspace ($\mathbf{U}_{i,\mathrm{M}}^{\perp}$, Extended Data Fig. 6c) that was optimized to capture trial-averaged variance across all experiments. We found considerable variability in optimal values of $\alpha$ across monkeys and task epochs (Extended Data Fig. 8e; larger values for monkey V and movement epoch). Across all datasets in both monkeys, we chose $\alpha_{opt}$ = 200 for fits in the decision epoch and $\alpha_{opt}$ = 50 for fits in the movement epoch in order to simplify comparisons between monkeys.

## Analysis of residual dynamics

The optimal hyper-parameters ($d_{opt}$, $l_{opt}$, and $\alpha_{opt}$ ), were used, in one final step, to estimate the time-varying dynamics matrices $\mathbf{A}_t$ (Equation 27) using all trials in $\tilde{\mathbf{Z}}$, separately for each choice, task epoch, and task configuration. We analyzed the resulting eigenvalue and singular value spectra of $\mathbf{A}_t$ at all times *t* in the trial. The eigenvalues/eigenvectors at the very first time-step were sorted in descending order of their eigenvalue magnitudes. At subsequent times, we sorted eigenvalues and the associated eigenvectors such that they were maximally consistent with those at the preceding time-step, using a modified version of an open-source MATLAB script (eigenshuffle.m[75]). A similar procedure was used to sort the time-varying singular values and the associated right and left singular vectors.

We computed the time constant of the dynamics (Fig. 4b,e) directly from the eigenvalue magnitudes of $\mathbf{A}_t$ as follows:

$$\tau_t^j = \frac{\Delta t}{\log(|\lambda_t^j|)} \tag{28}$$

where, $\lambda_t^j$ is the eigenvalue at time *t* associated with the *j^th^* eigenmode, and Δt is the time step (= 45 ms, length of time bin).

We analyzed the imaginary components of the complex-valued eigenvalues of $\mathbf{A}_t$ to obtain evidence for rotational dynamics (Fig. 4d,g). The natural oscillation frequency associated with the $j^{th}$ eigenmode is given by:

$$f_t^j = \frac{\angle\lambda_t^j}{2\pi\Delta t} \tag{29}$$

where $\angle\lambda_t^j$ is the angular phase of the $j^{th}$ eigenvalue.

We computed the largest eigenvalue magnitude ($|\lambda_t^{max}|$, Data Fig. 4e) and singular value ($|\sigma_t^{max}|$, Data Fig. 4f) at time *t* as:

$$\begin{aligned} |\lambda_t^{max}| &= \max_j |\lambda_t^j| \\ |\sigma_t^{max}| &= \max_j |\sigma_t^j| \end{aligned} \tag{30}$$

We quantified the magnitude of non-normality of the dynamics (Fig. 4h) based on a previously proposed measure[76] , which compared the singular values and eigenvalues magnitudes as follows:

$$d_F(\mathbf{A}_t) = \frac{\sqrt{\sum_j (\sigma_t^j)^2 - \sum_j (|\lambda_t^j|)^2}}{\sqrt{\sum_j (|\lambda_t^j|)^2}} \tag{31}$$

where, $\sigma_t^j$ and $\lambda_t^j$ are the $j^{th}$ singular and eigenvalue respectively.

## Task activity subspaces

### *Computing average task activity subspaces*

We used the 20-dimensional, aligned, single trial response patterns ($\mathbf{Z}$, output of Step 1 in Extended Data Fig. 2), to compute four distinct task activity subspaces. These four subspaces (denoted $\mathbf{U}_{task}^j$, $j$ ∈{choice, time, $\mathrm{jPC}_{12}$, $\mathrm{jPC}_{34}$}) captured variance in the aligned, trial-averaged trajectories due to choice (condition-dependent), time (condition-independent), and rotations[36] (Fig. 3c-d) and were computed separately for the decision (aligned to dots onset) and movement (aligned to movement onset) epochs, and for each task configuration.

To compute “choice” and “time” subspaces, trials in $\mathbf{Z}^i$ (*i* indexes experiments) were assigned to one of two choice conditions (choice 1 or choice 2), pooled across all experiments (within a task configuration) and then averaged, resulting in two trial-averaged response matrices $\langle\mathbf{Z}\rangle_{\text{choice}=1}$ and $\langle\mathbf{Z}\rangle_{\text{choice}=2}$ of dimensionality $20 \times \mathrm{T}_{epoch}$ ($\mathrm{T}_{epoch}$ = number of time bins in a single task epoch). We then compute a normalized "difference response matrix" ($\boldsymbol{D}$) and a "sum response matrix" ($\boldsymbol{S}$) as follows:

$$\begin{aligned}\boldsymbol{D} &= 0.5 * (\langle \mathbf{Z} \rangle_{\text{choice=1}} - \langle \mathbf{Z} \rangle_{\text{choice=2}}) \\ \boldsymbol{S} &= 0.5 * (\langle \mathbf{Z} \rangle_{\text{choice=1}} + \langle \mathbf{Z} \rangle_{\text{choice=2}})\end{aligned} \tag{32}$$

The first two principal components of the difference response matrix ($\boldsymbol{D}$) together defined the "choice" subspace, and captured most of the variance in response patterns due to *differences* between choices. Similarly, the first two principal components of the sum response matrix ($\boldsymbol{S}$) together constituted the "time" subspace, capturing maximal variance due to choice-independent components of aligned activity patterns.

To compute the jPC subspaces, we temporally smoothed single trials in $\mathbf{Z}^i$ (box filter, width = 180ms) before computing the trial-averaged response matrices $\langle \mathbf{Z} \rangle_{\text{choice=1}}$ and $\langle \mathbf{Z} \rangle_{\text{choice=2}}$ as described previously. The jPC vectors for the decision and movement epochs were estimated using these trial-averaged responses restricted to narrow time windows in each epoch ([500 1000]ms aligned to dots onset, as evidence for rotational dynamics was strongest at these times in the decision epoch, Fig. 3c,4d; [-250 250]ms aligned to movement onset, as rotational dynamics could underly movement related responses[36], also see Fig. 3d). jPC vectors were computed in the space spanned by the top 4 principal components (computed jointly on $\langle \mathbf{Z} \rangle_{\text{choice=1}}$ and $\langle \mathbf{Z} \rangle_{\text{choice=2}}$), without removing the condition-independent components of neural activity[36], resulting in two orthogonal jPC planes (jPC$_{12}$ and jPC$_{34}$, Fig. 3c-d), each spanned by a pair of complex-conjugate jPC vectors ($\mathbf{v}_1$ and $\mathbf{v}_2$). To determine the projection of the responses onto a single jPC subspace, we computed a pair of normalized real-valued vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ as $\mathbf{u}_1 = \mathbf{v}_1 + \mathbf{v}_2$ and $\mathbf{u}_2 = \mathrm{j} * (\mathbf{v}_1 - \mathbf{v}_2)$, which spanned the same subspace as $\mathbf{v}_1$ and $\mathbf{v}_2$. The imaginary components of the eigenvalues associated with $\mathbf{v}_1$ and $\mathbf{v}_2$ specified the natural frequency of rotation associated with a jPC plane. The jPC planes were ordered in descending order of their associated rotation frequency.

Only the two jPC planes (jPC$_{12}$ and jPC$_{34}$) were constrained to be mutually orthogonal (see Supplementary Analyses for alignment between other task activity subspace pairs). The task-activity subspaces $\mathbf{U}_{task}^{j}$ capture variance in the aligned, trial-averaged trajectories, but, need not perfectly align with the 8-dimensional dynamics subspace $\mathbf{U}_{dyn}^{d_{opt}}$ computed using the residuals. To assess the extent of the overlap between these two subspaces, we computed the fraction of total variance in a given task activity subspace that was attributable to activity unfolding within the dynamics subspace as follows:

$$\frac{\mathrm{Tr}\left(\mathrm{Cov}\left(\mathbf{U}_{dyn}^{d_{opt}}\left(\mathbf{U}_{dyn}^{d_{opt}\,\prime}\,\mathbf{U}_{task}^{j}\right)\mathbf{U}_{task}^{j\;\prime}\langle\mathbf{Z}\rangle\right)\right)}{\mathrm{Tr}\left(\mathrm{Cov}\left(\mathbf{U}_{task}^{j}\mathbf{U}_{task}^{j\;\prime}\langle\mathbf{Z}\rangle\right)\right)} \tag{33}$$

where $\mathrm{Tr}(.)$ is the matrix trace operator, $\mathrm{Cov}(.)$ corresponds to the covariance matrix of the argument, and $\langle\mathbf{Z}\rangle$ is the matrix of aligned, condition-averaged trajectories of size $20 \times (\mathrm{T}_{epoch} \times 2)$. We computed a null distribution by replacing the numerator of Equation 33 by $\mathrm{Tr}\left(\mathrm{Cov}\left(\mathbf{U}_{dyn}^{d_{opt}}\left(\mathbf{U}_{dyn}^{d_{opt}\,\prime}\,\mathbf{U}_{rand}\right)\mathbf{U}_{task}^{j\;\prime}\langle\mathbf{Z}\rangle\right)\right)$, where $\mathbf{U}_{rand}$ (sampled randomly 5000 times) is a pair of a random, orthogonal directions embedded within the 20-dimensional, aligned space. The resulting null distribution provides the range of possible values for the above fraction that could occur due to chance alignment of the 8-dimensional dynamics subspace with an arbitrary two dimensional subspace embedded within the 20-dimensional aligned space. The fraction of variance explained was in the range 0.66-0.94 (median = 0.85, n = 32, 2 task epochs x 4 planes x 4 task configurations) for monkey T, with 31/32 (32/32) datapoints lying beyond the 99th (95th) percentile of the null distribution. The range was 0.41-0.95 (median = 0.72, n =32) for monkey V, with 25/32 (28/32) datapoints beyond the 99th (95th) percentile of the null distribution. These findings imply that the components of the dynamics revealed by projections onto the task activity subspaces largely and consistently unfold within the dynamics subspace estimated using the residuals.

### *Comparison of residual eigenvectors to task activity subspaces*

We computed the alignment between each task activity subspace ($\mathbf{U}_{task}^{j}$) and the eigenvectors of the residual dynamics, separately within each task epoch. For each real-valued eigenvalue (pooled across 8 dimensions, times within epoch, and choices), we computed the subspace angle between a chosen task activity subspace and the associated real-valued eigenvector. For every estimated complex-conjugate eigenvalue pair, we computed a pair of subspace angles between a 2-dimensional eigenplane spanned by a pair of real-valued projection vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ (computed as described previously, i.e $\mathbf{u}_1 = \mathbf{v}_1 + \mathbf{v}_2$ and $\mathbf{u}_2 = \mathrm{j} * (\mathbf{v}_1 - \mathbf{v}_2)$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the complex-conjugate eigenvector pair) and the task activity subspace. To compute these subspace angles, we projected each eigenvector/eigenplane through the columns of $\mathbf{U}_{dyn}^{d_{opt}}$, back into the 20-dimensional aligned space to ensure that they were of the same dimension as vectors defining $\mathbf{U}_{task}^{j}$. We quantified the relationship between the eigenvalues and the alignment of the corresponding eigenvector/eigenplanes with the task activity subspace using a linear model (Fig. 5e). Specifically, we regressed each subspace angle indexed by a given task activity subspace (y-axis, Fig. 5e) against the corresponding eigenvalue magnitude (|ev|, x-axis in Fig. 5e, top) and rotation frequency (freq, x-axis in Fig. 5e, bottom; Equation 29) as shown below:

$$alignment(EV, j) = \beta_1^j |ev| + \beta_2^j freq \tag{34}$$

where, $j$ indexes the individual task -activity subspaces ($j$ ∈ {choice, time, jPC$_{12}$, jPC$_{34}$}). The regression coefficients were estimated using least squares and we also reported the 95% confidence intervals (based on the t-statistic) for each regression coefficient (errorbars in Fig. 5f).

## Robustness of the analysis pipeline

### *Estimating bias in the estimates of residual dynamics*

The choice of bin-size for binning spike-counts was critical for avoiding biases in estimates of $\mathbf{A}_t$ (Extended Data Fig. 9). To illustrate the effect of bin-size on the quality of estimates, we simulated data from a continuous-time, *time-invariant* linear dynamical system with a linear gaussian observation model:

$$\begin{aligned} \dot{\mathbf{x}}_t &= \mathbf{A}\mathbf{x}_t + \mathbf{b} + \boldsymbol{\epsilon}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{d} + \boldsymbol{\eta}_t \\ \mathbf{x}_1 &\sim \mathcal{N}(0, \mathbf{Q}_0) \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(0, \mathbf{Q}) \\ \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \tag{35}$$

where, $\mathcal{N}(.,.)$ denotes a normal distribution. We simulated 5000 single-trial trajectories for a total of 1500 time steps (1ms time steps) from a system with 3 latent dimensions and 20 observed dimensions. The elements of the three eigenvectors of $\mathbf{A}$ were sampled randomly from a standard normal distribution, and were orthogonalized (normal dynamics), and normalized to unit norm. The three eigenvalues were set to (-2, -4, -6) indicating stable, strongly decaying dynamics. The input vector $\mathbf{b}$ was set to $[2\ \ 2\ \ 2]^{\mathrm{T}}$. The covariance of the latent noise ($\boldsymbol{\epsilon}_t$) was set to a scaled identity matrix ($\mathbf{Q} = \sigma^2\mathbf{I}$). The values of $\sigma^2$ (grey lines, Extended Data Fig. 9a) were swept across two orders of magnitude to assess how latent noise variance affects estimates of $\mathbf{A}_t$. The observation matrix $\mathbf{C}$ was a random (elements sampled from a standard normal distribution), orthogonal matrix**.** The elements of the baseline input vector $\mathbf{d}$ were sampled from a uniform distribution between $[0,8]$. The observation noise matrix $\mathbf{R}$ was diagonal, with elements sampled from a uniform distribution between $[0,0.05]$. The initial noise covariance ($\mathbf{Q}_0$) was obtained by solving the continuous time Lyapunov equation.

We estimated the dynamics matrix ($\mathbf{A}_t$) using residuals binned in non-overlapping bins of sizes [2, 3, 5, 10, 15, 30, 40, 60]ms. Specifically, we chose a hankel order $q$ = 5 (Equation S14, Supplementary Math Note A) and did not optimize hyper-parameters $l$ and $\alpha$ during estimation.

Instead, we chose $l$ and $\alpha$ sensibly, ensuring that they were consistent across different bin sizes and with the underlying model class. For instance, $\alpha$ was set to a large value (= $10^6$) ensuring time-invariant fits, while $l$ was set such that it roughly translated into equal units of time for different bin sizes.

We assessed the effect of bin-size on the estimated eigenvalues of $\mathbf{A}_t$. Importantly, eigenvalues for different bin-sizes cannot be compared directly, as discretizing a continuous-time dynamical system trivially results in eigenvalues that depend on the duration of the discretization time-step (here the bin size). The same dynamical system, when expressed at step-sizes $\Delta\mathrm{t_j}$ and $\Delta\mathrm{t_{ref}}$, would therefore result in eigenvalues $\hat{\lambda}_{\Delta\mathrm{t_j}}$ and $\hat{\lambda}^{(\mathrm{j})}_{\Delta\mathrm{t_{ref}}}$ related by the following scaling relation:

$$\hat{\lambda}^{(\mathrm{j})}_{\Delta\mathrm{t_{ref}}} = (\hat{\lambda}_{\Delta\mathrm{t_j}})^{\frac{\Delta t_{ref}}{\Delta t_j}} \tag{36}$$

To discount these trivial differences, we transformed each estimated eigenvalue $\hat{\lambda}_{\Delta\mathrm{t_j}}$ obtained for bin size $\Delta\mathrm{t_j}$ into an 're-binned' eigenvalue $\hat{\lambda}^{(\mathrm{j})}_{\Delta\mathrm{t_{ref}}}$ expected for a reference bin size $\Delta\mathrm{t_{ref}}$ = 40ms, and compared them to the 'ground truth' eigenvalue expected for a bin size of $\Delta\mathrm{t_{ref}}$ (Extended Data Fig. 9a). The absolute value of $\hat{\lambda}^{(\mathrm{j})}_{\Delta\mathrm{t_{ref}}}$ asymptotically converged to the ground truth for increasing bin sizes, meaning that large bin sizes resulted in unbiased estimates, with convergence being independent of the specific choice of $\Delta\mathrm{t_{ref}}$.

We observed a similar asymptotic convergence for the neural data (Extended Data Fig. 9b). This observation was used to determine the optimal bin size for which estimates of residual dynamics can be expected to be unbiased. We binned the recorded spiking data for monkey T in bin sizes of [15, 30, 45, 60, 90]ms and projected the resulting single trial trajectories (for all bin sizes) into a common aligned subspace (Step 1, Extended Data Fig. 2; $\mathbf{U}^{\perp}_{i,\mathrm{M}}$ , Equation 14) determined for a bin size of 45ms, before computing residual dynamics. Once again, we did not optimize the hyper-parameters ($d$, $l$, and $\alpha$; Steps 3 and 4, Extended Data Fig. 2) of the pipeline, as the aim was to understand how bin size alone effects the estimated eigenvalues. Instead, we fixed values of $d$ and $\alpha$ to the optimal ones determined by cross-validation (for residuals binned in 45ms bins) described previously (Extended Data Fig. 8, Fig. 4; $d$ = 8, $\alpha$ = 200/50 for decision/movement epochs) Values of $l$ were instead chosen separately for each bin size such that it roughly translated into equal units of time ($l$ = 3/2 for bin sizes of 45/60ms implying a 135/120ms long window in the past, Equations 22-24).

We computed the 're-binned' eigenvalue magnitudes of $\mathbf{A}_t$ (Equation 36) for the different bin sizes, expected under a reference bin size ($\Delta\mathrm{t_{ref}}$) of 15ms. The 're-binned' eigenvalue corresponding to each of the 8 eigen-modes was averaged across time within two distinct time windows that exhibited the most pronounced temporal dependencies (Fig. 4; $t \in [200\ 400]$ms aligned to dots onset, and $t \in [-150\ 250]$ms aligned to movement onset). We observed asymptotic convergence for all 8 eigen-modes in all task conditions (choice 1 or choice 2), task epochs (decision or movement), and all task configurations in monkey T for bin sizes greater than 30ms (Extended Data Fig. 9b). Based on these findings, a bin size of 45ms was well motivated for our analyses.

### *Qualitative estimates of goodness of fit*

The average, cross-validated mean squared error (computed using held-out test trials) of the predictions resulting from the first stage of 2SLS ($E_{fs}$, Equation 26; Extended Data Fig. 8b, shown only for a single configuration in monkey T) for optimal values of hyperparameters ($d_{opt}$ = 8, $l_{opt}$=3) translated into a coefficient of determination ($R^2$) of 0.0367/0.0390 (mean across all task configurations and choices, std = 0.0064/0.0029) in monkey T/V respectively. Similar $R^2$ values (mean(std) = 0.0577(0.014)/0.065(0.013) for monkey T/V) were obtained for predictions resulting from the second stage of 2SLS (Extended Data Fig. 8b, shown only for a single configuration in monkey T) for optimal values of $\alpha_{opt}$ (=200/50 for the decision/movement epochs)

The small $R^2$ values made it difficult to gauge 'goodness of fit', as their relatively small magnitudes could be due to unstructured observation noise (which cannot be predicted by any model) dominating the variability in the residuals. To determine if this is indeed true, we simulated residuals from a time-varying linear dynamical system (Equation 11) with dynamics matrix ($\mathbf{A}_t$) and the observation matrix ($\mathbf{C}$) matched to optimal estimates of residual dynamics and dynamics subspace ($\mathbf{U}_{dyn}^{d_{opt}}$) obtained using neural data.

Simulating observed residuals ($\tilde{\mathbf{z}}_t$, Equation 11) also required, as a first step, estimating the latent noise covariance ($\mathrm{Cov}(\boldsymbol{\epsilon}_t) = \mathbf{Q}$), variance in the initial latent state ($\mathrm{Cov}(\tilde{\mathbf{x}}_0) = \mathbf{Q}_0$), and the observation noise covariance ($\mathrm{Cov}(\boldsymbol{\eta}_t) = \mathbf{R}$ ). Closed form estimates for these parameters were obtained using maximum likelihood:

$$\hat{\mathbf{R}} = \frac{1}{(T-l).K} \mathrm{diag}\Big( \sum_{t=l+1}^{T} \sum_{k=1}^{K} [\, \tilde{\mathbf{z}}_t(k) - \mathbf{U}_{dyn}^{d_{opt}} \hat{\tilde{\mathbf{x}}}_t(k)] [\tilde{\mathbf{z}}_t(k) - \mathbf{U}_{dyn}^{d} \hat{\tilde{\mathbf{x}}}_t(k)]' \Big) \tag{37}$$

$$\hat{\mathbf{Q}}_0 = \frac{1}{\mathrm{K}} \sum_{k=1}^{K} \hat{\tilde{\mathbf{x}}}_{l+1}(k) \hat{\tilde{\mathbf{x}}}_{l+1}(k)' \tag{38}$$

$$\widehat{\mathbf{Q}} = \frac{1}{d.(T-l-1).K} Tr\left(\sum_{t=l+1}^{T-1}\sum_{k=1}^{K}[\hat{\tilde{\mathbf{x}}}_{t+1}(k) - \mathbf{A}_t\hat{\tilde{\mathbf{x}}}_t(k)][\hat{\tilde{\mathbf{x}}}_{t+1}(k) - \mathbf{A}_t\hat{\tilde{\mathbf{x}}}_t(k)]'\right) \tag{39}$$

where, $\hat{\tilde{\mathbf{x}}}_t(k)$ is the denoised prediction of the residual latent state at time *t* on trial *k* (Equation 24) resulting from the first stage of 2SLS.

These estimates were used to simulate residuals (Equation 11) for a matched number of trials for each choice and task configuration, which were then used to compute idealized coefficients of determination ($R^2_{\text{sim-fs}}$ and $R^2_{\text{sim-ss}}$), under the assumption that our analysis pipeline works perfectly i.e. is able to perfectly retrieve the dynamics (second stage of 2SLS) and the de-noised residual latent states (first stage of 2SLS) at each time. We reasoned that this would provide a realistic benchmark, if not a strict upper limit, for the fit quality that one can hope to obtain in the context of large observation noise.

To compute $R^2_{\text{sim-fs}}$ and $R^2_{\text{sim-ss}}$, we projected the simulated residual observations (${\tilde{\mathbf{z}}_t}^{sim}(k)$) into the estimated dynamics subspace ($\mathbf{U}_{dyn}^{d_{opt}}$, Equation 21) and computed the amount of variance explained in the resulting projection by (i) the simulated latent state (denoted by ${\tilde{\mathbf{x}}_t}^{sim}(k)$), and by (ii) a 'noise-free', one-step propagation of the simulated latent state through the corresponding estimate of the dynamics matrix. The former ($R^2_{\text{sim-fs}}$) provides a benchmark for comparing the coefficient of determination obtained for the first stage of the 2SLS, whereas the latter ($R^2_{\text{sim-ss}}$) provides the same for the second stage of the 2SLS. Mathematically, these quantities were defined as follows:

$$\mathrm{R}^2_{sim-fs} = 1.0 - \frac{\sum_t \sum_k \left(\mathbf{U}_{dyn}^{d_{opt}\,'} . {\tilde{\mathbf{z}}_t}^{sim}(k) - {\tilde{\mathbf{x}}_t}^{sim}(k)\right)^2}{\sum_t \sum_k \left(\mathbf{U}_{dyn}^{d_{opt}\,'} {\tilde{\mathbf{z}}_t}^{sim}(k)\right)^2} \tag{40}$$

$$\mathrm{R}^2_{sim-ss} = 1.0 - \frac{\sum_t \sum_k \left(\mathbf{U}_{dyn}^{d_{opt}\,'} . {\tilde{\mathbf{z}}_t}^{sim}(k) - \mathbf{A}_{\mathrm{t}-1}{\tilde{\mathbf{x}}_{t-1}}^{sim}(k)\right)^2}{\sum_t \sum_k \left(\mathbf{U}_{dyn}^{d_{opt}\,'} {\tilde{\mathbf{z}}_t}^{sim}(k)\right)^2} \tag{41}$$

where, ${\tilde{\mathbf{x}}_t}^{sim}(k) = \mathbf{A}_{t-1}{\tilde{\mathbf{x}}_{t-1}}^{sim}(k) + \hat{\boldsymbol{\epsilon}}_{t-1}$, and $\hat{\boldsymbol{\epsilon}}_{t-1}$ is a sample from a multi-variate gaussian with covariance $\widehat{\mathbf{Q}}$.

The range of values of $R^2_{\text{sim-fs}}$ (monkey T : 0.0738 ± 0.011, monkey V : 0.0958 ± 0.33; mean ± std across task configurations and choices) and $R^2_{\text{sim-ss}}$ (monkey T : 0.0512 ± 0.0171, monkey V:

0.0674 ± 0.0157; mean ± std) qualitatively matched the range of values of the corresponding cross-validated coefficient of determination ($R^2$) for the first and second stage of 2SLS obtained for the data (reported above). This finding implies that the low coefficient of determination measured in the real data is likely due to residuals being dominated by uncorrelated observation noise.

## Simulated Models

We validated our analysis pipeline on a number of simulated models, which were broadly categorized into four groups – (1) models of decisions and movement, (2) linear state-space models with uncorrelated latent noise, (3) linear state-space models with correlated latent noise, (4) modular two area recurrent network model. The first two model categories exemplified the simple input regime (Fig. 1b), whereas the latter two represented the complex input regime (Fig. 1b). We only provide a brief description of these four model categories here (see Supplementary Methods for details).

### *Models of decisions and movement*

We simulated single trial responses from 6 distinct models; three of these corresponded to "models of decisions" (saddle point[35], line attractor[12], and point attractor), and the other three to "models of movement" (rotational dynamics[36], dynamic attractor[37], and point attractor). Within each sub-category (decision or movement), the three models had distinct recurrent dynamics and time-varying input drives, informed by previous models of sensory evidence integration and movement generation, but defined so as to exhibit the same condition average trajectories (Fig. 1c-d, Supplementary Methods). All six models were described by a 2-dimensional latent state ($\mathbf{x}$) governed by Equation 1 (see Supplementary Methods for specifications of parameters in Equation 42). Observed states ($\mathbf{y}$) resulted from a linear gaussian observation process (similar to Equation 11, but with $\mathbf{C} = \mathbf{I}$) as defined below:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{F}(\mathbf{x}) + \mathbf{u}_t + \boldsymbol{\epsilon}_t \\ \mathbf{y}_t &= \mathbf{x}_t + \boldsymbol{\eta}_t\end{aligned} \tag{42}$$

For each model, we simulated a total of 4000 trials, each of duration of 1s (steps of 1ms). Each trial belonged to one of two conditions (choice 1 or choice 2) determined either by the initial condition of the recurrent dynamics or the inputs. We estimated the time-varying residual dynamics ($\mathbf{A}_t$, Equation 11) using only the 2SLS regression (Steps 3 and 4 of analysis pipeline, Extended Data Fig. 2) directly on the two-dimensional residuals (without Steps 1 and 2 of the analysis pipeline, Extended Data Fig. 2). We did not optimize any of the hyper-parameters of the pipeline. We used dimensionality ($d$) = 2, lag ($l$) = 5 and a regularization parameter ($\alpha$) = 100 for for all model fits.

To illustrate the various hypothesized relationships between the residual dynamics and the condition-averaged trajectories (Fig. 5), we also simulated an “augmented” line attractor model (Fig. 5a-b) and an “augmented” rotational dynamics model (Fig. 5c-d), both characterized by two additional latent dimensions (4 latent dimensions in total). The two additional latent dimensions were orthogonal to the first two latent dimensions and were associated with quick, decaying dynamics and sinusoidal inputs (Supplementary Methods).

### *Linear state-space models with uncorrelated latent noise*

We also validated the analysis pipeline on simulated single trial responses from six distinct latent-variable state space models (Extended Data Fig. 3), which were characterized by (i) three distinct *linear but time-varying* latent dynamics (Equation 11, also see Supplementary Methods), and (ii) two distinct observation models: linear-gaussian (Equation 11) or poisson (Supplementary Methods). Additionally, we simulated three more models characterized by linear-gaussian observations (Equation 11) but subject to both time-varying dynamics *and* time-varying latent noise (Supplementary Methods). These simulations demonstrated the robustness of our pipeline to different latent dynamics and observation model types. For all nine models, residuals were binned in 45ms bins (poisson observations were square-root transformed), and were subjected to steps 2-4 of the analysis pipeline (Extended Data Fig. 2, without session alignment), using cross-validation to tune the hyper-parameters.

### *Linear state-space models with correlated latent noise*

To study the inflationary effects of correlated, latent input noise ($\xi$(t) in Fig. 1b, complex input) on estimates of residual dynamics , we considered state-space models with linear time-invariant dynamics, characterized by latent noise with decaying temporal autocorrelations (correlated noise). We used these models to understand how neural activity that is a consequence of recurrent processing in unobserved/unrecorded areas influences residual dynamics measured within recorded/observed areas. To model correlated latent noise, we assumed a time-invariant, linear state-space model governed by the following set of equations.

$$\begin{aligned} \mathbf{x}(\mathrm{t}+1) &= \mathbf{A}\mathbf{x}(\mathrm{t}) + \boldsymbol{\varepsilon}(\mathrm{t}) \\ \boldsymbol{\varepsilon}(\mathrm{t}+1) &= \boldsymbol{\phi}\,\boldsymbol{\varepsilon}(\mathrm{t}) + \boldsymbol{\zeta}(\mathrm{t}) \end{aligned} \tag{43}$$

where, $\boldsymbol{\zeta}(t)$ is a zero-mean white gaussian noise process with covariance matrix $\mathbf{Q}$. We interpreted the model specified in Equation 43 as follows. $\mathbf{x}(\mathrm{t})$ was assumed to represent the latent population state of the recorded/observed area, yielding observations $\mathbf{z}(\mathrm{t})$. $\boldsymbol{\varepsilon}(\mathrm{t})$ was considered to represent the latent population state of unobserved or unrecorded areas, contributing auto-correlated, latent input noise process (similar to $\boldsymbol{\xi}_t$, Fig. 1b, complex input) that directly influenced $\mathbf{x}(\mathrm{t})$. Therefore $\boldsymbol{\phi}$**,** which determined the dynamics of $\boldsymbol{\varepsilon}(\mathrm{t})$, reflected the dynamics of the unobserved/unrecorded areas. For a given $\mathbf{A}$, $\boldsymbol{\phi}$ and $\mathbf{Q}$, we were able to analytically derive the two-stage least squares (2SLS) estimate of the residual dynamics (see Supplementary Math Note B), assuming that (i) the model operated in *steady-state* and, (ii) that we only had access to $\mathbf{x}(\mathrm{t})$. We systematically varied $\mathbf{A}$, $\boldsymbol{\phi}$ in order to quantify the effect of correlated, latent input noise on analytically derived estimates of residual dynamics, under steady-state conditions (Extended Data. Fig. 4,5; Supplementary Methods, Supplementary Math Note B). These analyses did not require specifying an observation model (unlike the previous models).

### *Modular two-area recurrent neural networks*

We simulated single trial responses using a modular, two area RNN model of perceptual decision making, which emulated the interactions between posterior parietal cortex (PPC) and pre-frontal cortex (PFC)[38]. Each area was characterized by two choice-selective (choice 1 or choice 2) neural populations, which were recurrently interconnected through excitatory and inhibitory (E/I) intra-areal (within area) connections. These neural populations were also interconnected across areas though inter-areal (between areas), E/I, feedforward and feedback connections.

We denote the state of area *a* (local state) at time *t* as $\mathbf{x}_t^a$, a 2-dimensional vector (one dimension per choice selective population in area *a*). The 'global' network state $\mathbf{x}_t$ (4 dimensional) was defined by concatenating the local state across both areas (Equation 44). Observations specific to area *a*, denoted by $\mathbf{y}_t^a$ (10-dimensional), were obtained through a linear-gaussian observation model ($\boldsymbol{\eta}_\mathrm{t}$ is multivariate, isotropic gaussian, with variance equal to 0.0006) applied to the 'global' state. The observation matrix ($\mathbf{C}_{model}$) was block-diagonal (each block representing the observation matrix specific to an area):

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{y}_t^{ppc} \\ \mathbf{y}_t^{pfc} \end{pmatrix} = \mathbf{C}_{\mathrm{model}} \mathbf{x}_t + \boldsymbol{\eta}_t = \begin{bmatrix} \mathbf{C}_{\mathrm{ppc}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathrm{pfc}} \end{bmatrix} \begin{pmatrix} \mathbf{x}_t^{ppc} \\ \mathbf{x}_t^{pfc} \end{pmatrix} + \boldsymbol{\eta}_t \tag{44}$$

Considering that each area is characterized by two choice-selective populations, the task relevant dimensions corresponding to "choice" and "time" modes specific to each area (Fig. 6a,d) were naturally defined in the 4-dimensional 'global' stat space space as:

$$\begin{aligned}\boldsymbol{u}_{choice}^{ppc} &= [1\ \ -1\ \ 0\ \ 0]^T\\ \boldsymbol{u}_{time}^{ppc} &= [1\ \ 1\ \ 0\ \ 0]^T\\ \boldsymbol{u}_{choice}^{pfc} &= [0\ \ 0\ \ 1 - 1]^T\\ \boldsymbol{u}_{choice}^{pfc} &= [0\ \ 0\ \ 1 - 1]^T\end{aligned} \tag{45}$$

We simulated two different types of networks, one in which feedback from PFC to PPC was absent and another in which feedback was present (see details in Supplementary Methods). For each network type, we simulated thirty different network configurations with distinct intra-areal and inter-areal connectivity strengths, parameterized using scalar-valued parameters $\mathrm{J}_{self}$ (5 distinct values; colored markers, Fig. 6c,f) and $\mathrm{J}_{across}$ (6 distinct values; x-axis Fig. 6c,f) respectively. For the network configuration shown in Fig. 6a ($\mathrm{J}_{self}$ = 0.36, $\mathrm{J}_{across} = 0.08$, no feedback), we simulated an identical network (with frozen noise) for a 'shuffled' condition, in which, only the feed-forward current inputs at each time from PPC to PFC were randomly shuffled across trials, in order to remove any slow temporal autocorrelations (Fig. 6b). Only PPC was driven using external input on each trial (indexed by *k*) defined as follows:

$$I^k(t) = \begin{cases} 0\,, & 0 < t \le T_{on} \\ I_e\left(1 \pm \dfrac{c(k)}{100\%}\right), & t > T_{on} \end{cases} \tag{46}$$

where, $I_e$ = 0.0130nA, $T_{on}$ (= 400ms) is the time of stimulus onset and, *c(k)* corresponds to the coherency on the *k*th trial. We simulated only trials with zero coherency (c(k) = 0) and assigned each trial as either "choice-1" or "choice-2", depending on the population 'choice' readout from PFC (projection onto $\boldsymbol{u}_{choice}^{pfc}$) at the last time step of the trial. Specific details about network architecture and dynamics can be found in ref[38].

Residual dynamics was estimated either 'locally' (see Fig. 6b), using observations of PPC alone ($\mathbf{y}_t^{ppc}$) or PFC alone ($\mathbf{y}_t^{pfc}$); or 'globally' (Fig. 7a), using observations from both areas ($\mathbf{y}_t$). Observations were temporally binned in 45ms long bins and residual dynamics was computed separately for each choice condition by employing the full analysis pipeline (Supplementary Methods), but excluding the session alignment (step 1 in Extended Data Fig. 2). Additionally, we also computed the *'local choice'* residual dynamics, by fitting the 1-dimensional projection of residuals in PPC and PFC onto their respective choice dimensions, $\boldsymbol{u}_{choice}^{ppc}$ and $\boldsymbol{u}_{choice}^{pfc}$. We examined the relationship between the largest eigenvalue magnitude (across time in the trial) of the *'local choice'* residual dynamics (y-axis in Fig. 6c,f ; error bars are 95% bootstrap confidence intervals) and the network connectivity parameters $J_{self}$ (colors in Fig. 6c,f) and $J_{across}$ (x-axis in Fig. 6c,f).

We performed a set of targeted causal perturbation experiments (Fig. 8, Extended Data Fig. 10) for the two example network configurations (Fig. 6 & 7). We first obtained a set of "ground truths" that summarized how activity patterns associated with each area change in response to a simulated perturbation. We then compared the simulated perturbations to predictions based on either the 'local' or 'global' estimates of residual dynamics (Supplementary Methods).

## Methods-Only References

67. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).

68. Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A. & Shenoy, K. V. A high-performance brain-computer interface. *Nature* **442**, 195–198 (2006).

69. Katayama, T. *Subspace Methods for System Identification*. (Springer London, 2005). doi:10.1007/1-84628-158-X.

70. Cedervall, M. & Stoica, P. System identification from noisy measurements by using instrumental variables and subspace fitting. *Circuits Syst. Signal Process.* **15**, 275–290 (1996).

71. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. (Springer, 2008).

72. Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Stat.* **11**, 284–300 (1940).

73. Bound, J., Jaeger, D. A. & Baker, R. M. Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *J. Am. Stat. Assoc.* **90**, 443–450 (1995).

74. Rudin, L. I., Osher, S. & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. Nonlinear Phenom.* **60**, 259–268 (1992).

75. D'Errico, J. Eigenshuffle, MATLAB Central File Exchange. (2009).

76. Henrici, P. Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Numer. Math.* **4**, 24–40 (1962).