Feedforward and feedback interactions between visual cortical areas use different population activity patterns

João D. Semedo[1], Anna I. Jasper[2], Amin Zandvakili[2], Amir Aschner[2],
Christian K. Machens[3†], Adam Kohn[2,4,5†], Byron M. Yu[1,6†*]

[†]These authors contributed equally to this work.

[*]Corresponding author.

[1]Electrical and Computer Engineering Dept., Carnegie Mellon University, Pittsburgh, PA, USA

[2]Dominick Purpura Dept. of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA

[3]Champalimaud Research, Champalimaud Centre for the Unknown, Lisbon, Portugal

[4]Ophthalmology and Visual Sciences Dept., Albert Einstein College of Medicine, Bronx, NY, USA

[5]Systems and Computational Biology Dept., Albert Einstein College of Medicine, Bronx, NY, USA

[6]Biomedical Engineering Dept., Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Brain function relies on the coordination of activity across multiple, recurrently connected, brain areas. For instance, sensory information encoded in early sensory areas is relayed to, and further processed by, higher cortical areas and then fed back. However, the way in which feedforward and feedback signaling interact with one another is incompletely understood. Here we investigate this question by leveraging simultaneous neuronal population recordings in early and midlevel visual areas (V1-V2 and V1-V4). Using a dimensionality reduction approach, we find that population interactions are feedforward-dominated shortly after stimulus onset and feedback-dominated during spontaneous activity. The population activity patterns most correlated across areas were distinct during feedforward- and feedback-dominated periods. These results suggest that feedforward and feedback signaling rely on separate "channels", such that feedback signaling does not directly affect activity that is fed forward.

Most brain functions rely on the coordination of activity across multiple areas[1,2]. Activity does not follow a purely feedforward path between brain areas: areas are often reciprocally connected, and signals passed from one area to the next are often processed and fed back[3–6]. Understanding when feedforward and feedback signaling between areas is most dominant, and how these forms of signaling interact, is crucial for improving our understanding of computation in the brain.

Previous studies have attempted to infer feedforward or feedback interactions between areas. One approach for identifying feedforward signaling is to present a stimulus and then compare the timing of neuronal response onsets across areas[7–10]. Similarly, feedback signaling can be inferred by studying time differences in the emergence of some forms of selectivity across areas[11–15]. Other studies have studied feedforward or feedback signaling by measuring activity simultaneously in two areas, and comparing temporal delays in pairwise spiking correlations[16–21] or phase delays in local field potentials (LFP)[22–25]. Most of these studies focused on the activity of pairs of neurons across areas, or aggregate measures of neural activity such as local field potentials.

To understand inter-areal interactions more deeply, it is now possible to record activity from large neuronal populations simultaneously in different cortical areas, and characterize what patterns of population activity are most related across those areas[19,26–33]. This approach has led to new proposals about how activity can be flexibly routed across brain areas (see ref. 34 for a review). In particular, simultaneous multi-area recordings have revealed properties of population activity patterns that are most related across areas in the context of sensory processing[29], attention[30], learning[31], and motor control[32,33]. However, it is unknown how these population activity patterns relate to feedforward or feedback signaling between areas.

Here, we leverage simultaneous recordings of neuronal populations in early and midlevel visual areas (V1-V2 and V1-V4) to examine the temporal dynamics of inter-areal interactions, as well as the population activity patterns involved in those interactions (Fig. 1a). We correlated the population activity across areas at different time delays to infer feedforward and feedback signaling. Interactions were feedforward-dominated (V1 leading V2, and V1 leading V4) shortly after stimulus onset and gradually became feedback-dominated with persistent stimulus drive, as well as during spontaneous activity. Importantly, the population activity patterns involved in feedforward signaling were distinct from those involved in feedback signaling. This indicates that activity patterns in V1 that most affect downstream activity during feedforward processing are not the ones most affected by feedback signaling, suggesting both forms of signaling can co-exist without interference. Our results reveal both the dominant direction of signal flow between areas on a moment-by-moment basis and the population activity patterns involved in feedforward and feedback interactions.

Results

We simultaneously recorded from neuronal populations in V1 (88 to 159 neurons; mean: 112.8 ± 12.3 SEM) and V2 (24 to 37 neurons; mean: 29.4 ± 2.4 SEM) in three anesthetized monkeys (Fig. 1b; five recording sessions), as well as in V1 (34 to 128 neurons; mean: 66.6 ± 16.2 SEM) and V4 (12 to 84 neurons; mean: 58.8 ± 12.4 SEM) in two awake fixating monkeys (Fig. 1c; five recording sessions). Animals were shown drifting gratings of different orientations (1280 ms stimulus duration for V1-V2; 200 ms for V1-V4), followed by a blank screen (1500 ms for V1-V2; 150 ms for V1-V4). Recording sites were chosen so that the spatial receptive fields of the V1 and V2/V4 populations overlapped (see ref. 19 and Supplementary Fig. 1).

Temporal structure of inter-areal interactions

We first characterized the temporal dynamics of the interaction between neuronal population spiking responses in V1 and V2. To do so, we asked: (1) how the interaction evolved during stimulus presentation and the subsequent period of spontaneous activity (which together constitute a trial); and (2) how the interaction depended on the time delay considered between the two areas. Given that these areas are reciprocally connected, with activity flowing in both directions, it is possible that there are periods during which V1 leads V2 activity, and other periods where it lags behind.

To measure interactions between areas, we employed Canonical Correlation Analysis (CCA). Consider representing the activity in two neuronal populations using two activity spaces, one for each area. In each space, each coordinate axis corresponds to the activity of a recorded neuron (Fig. 2a). Within a given time window, the spike counts of the neurons (in the two populations) define a point in each space. For each point in V1 activity space (Fig. 2a, left panel), there is a corresponding, simultaneously recorded point in V2 activity space (Fig. 2a, right panel). CCA seeks dimensions of activity in each area, such that activity along those dimensions is maximally correlated across the two areas (Fig. 2a, bottom panel). For this analysis, we focused on the most correlated dimensions across the two areas (i.e., the first canonical pair; correlations associated with the second canonical pair were on average 60% lower and close to chance level). We used the correlation value for the first canonical pair as a measure of inter-areal interaction strength, which we refer to as *population correlation*.

Interactions between areas likely involve time delays due to signal conduction, as well as network processing. This implies that the activity across areas might not be most related for matched (simultaneous) time windows, but for time windows shifted forward or backward in time. Thus, we used CCA to relate activity recorded in V1 with activity in V2 at different time delays (Fig. 2b; Methods) to produce a population correlation function (Fig. 2c). This population correlation function can be computed at different epochs in a trial.

We found that V1-V2 population correlations were lowest just after stimulus onset, increased steadily during stimulus presentation, and were highest for spontaneous activity (Fig. 3a). Focusing on the activity shortly after stimulus onset ("Early Evoked"; 160 ms after stimulus onset), population correlations were larger for positive delays than for negative delays (red trace in Fig. 3b, with peak correlation occuring for a lag of 3 ms), meaning V1 activity was most correlated with V2 activity occurring later in time –consistent with a feedforward interaction. The feedforward interaction became less evident later during the evoked activity period ("Late Evoked"; 1120 ms after stimulus onset; yellow trace in Fig. 3b). After stimulus offset, population correlations were larger for negative delays, so that V2 led V1, suggesting a feedback-dominated interaction ("Spontaneous", purple trace in Fig. 3b, with a broad peak centered at approximately -15 ms; 2240 ms after stimulus onset). For a more complete characterization, we show in Fig. 3c how population correlations vary as a function of time delay between areas (horizontal axis) and the time relative to stimulus onset (vertical axis; note that the population correlation functions in Fig. 3b represent horizontal slices of this representation).

To quantify the shift from feedforward- to feedback-dominated interactions, we calculated a feedforward ratio, defined as the difference between the feedforward (positive delay) and feedback (negative delay) sides of the population correlation function, divided by their sum. In every recording session, we found that V1-V2 interactions were more feedback-dominated during the spontaneous period than during the evoked period (Fig. 3d, left; average feedforward ratio, computed in the -80 to 80 ms delay range:

-0.005 ± 0.008 SEM for late evoked activity; -0.040 ± 0.011 SEM for spontaneous activity; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between late evoked and spontaneous activity across all 5 recording sessions; t-test for feedforward ratio, $p = 0.57$ for late evoked activity, $p = 0.02$ for spontaneous activity).

The population correlation functions contain both slow- and fast-timescale features. To isolate the fast-timescale features, particularly evident early in the evoked period (Fig. 3b, red), we computed jitter-corrected population correlation functions for responses measured after stimulus onset[35,36] and computed their peak location and height (Supplementary Fig. 2). Clear feedforward peaks will have large heights whereas the absence of a peak will result in a small peak height with highly variable peak times (i.e., reflecting "noise" in the correlation function). We found a clear, early feedforward peak in all recording sessions for which the V1 and V2 receptive fields were aligned (Fig. 3e, open circles; average peak height: 0.008 ± 0.002 SEM; average peak delay: 2.2ms ± 0.37 SEM).

If the effects shown in Fig. 3 truly reflect feedforward and feedback interactions, they should display appropriate retinotopic specificity. Feedforward connections are more retinotopically precise than feedback connections[37–41]. As a result, feedforward interactions should require retinotopic alignment, whereas feedback interactions might be more tolerant of retinotopic misalignment between the neurons sampled in the two areas. To test this prediction, we performed additional recordings for which the spatial receptive fields of the V1 and V2 populations were misaligned by several degrees (mean center-to-center population spatial receptive field distance was $3.73\,\mathrm{deg}$ for misaligned sessions and $0.58\,\mathrm{deg}$ for aligned sessions).

Population correlations were lower for these recordings than for those from populations with aligned receptive fields (Fig. 3a, dotted line). The fast time-scale correlation peaks observed shortly after stimulus onset for aligned populations (Fig. 3e, circles) were absent in responses from populations with misaligned receptive fields, evident as small peak heights and inconsistent peak delays (Fig. 3e, triangles; average peak height: 0.0025 ± 0.0001 SEM; one-sided permutation test, $p = 0.004$ for difference between sessions with aligned vs. misaligned receptive fields). Despite the absence of a clear feedforward peak, the V1-V2 interaction for the misaligned populations was still feedback-dominated during spontaneous activity (Fig. 3d, purple; average feedforward ratio: -0.003 ± 0.019 SEM for late evoked activity; -0.027 ± 0.013 SEM for spontaneous activity; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between late evoked and spontaneous activity across all 5 recording sessions). Thus, the feedforward and feedback interactions identified by CCA have properties consistent with the underlying anatomical specificity.

To test whether the dynamics of V1-V2 interactions might reflect in part changes in the activity within each area, rather than the interaction between areas, we devised two controls. First, we split each V1 and V2 population randomly into two groups, and measured within-area correlations as we had done when analyzing inter-areal interactions. The features described for inter-areal interactions were absent when identical analyses were performed on neurons recorded in the same area (Supplementary Fig. 3). Specifically, within-area interactions showed no evidence of a feedforward peak and were symmetric with respect to the time lag during late evoked and spontaneous activity. Thus, the changes in temporal structure shown in Fig. 3 are specific to inter-areal interactions.

Second, we tested whether the dynamics of inter-areal interactions might be related to differences in neuronal onset latency in the two areas, or to changes in the firing rates over time within each population. To assess this possibility, we performed CCA after shuffling the correspondence of trials in the two areas, while keeping the temporal correspondence within each area intact (see Methods). This shuffling procedure maintained the firing rate time courses and correlation structure within each area, but broke the trial-by-trial correspondence of activity across the two areas. After shuffling, inter-areal correlations no longer increased throughout the trial (Fig. 3a, light trace). Furthermore, there was no evidence of a feedforward interaction early in the trial, nor was there a shift to a feedback-dominated interaction during spontaneous activity (Fig. 3b, light traces). Thus, the dynamics of inter-areal interactions cannot be attributed to different onset latencies or response dynamics in the two areas.

We then asked whether inter-areal interactions showed similar dynamics in responses measured in awake animals as in the responses measured in anesthetized animals considered thus far. We recorded V1 and V4 population activity, in two animals performing a passive fixation task in which drifting gratings were presented (Methods). As with V1-V2 responses, V1-V4 population correlation increased throughout the evoked period (Fig. 4a; compare with Fig. 3a). Just after stimulus onset, V1-V4 interactions were feedforward-dominated (Fig. 4b, red curve; 75 ms after stimulus onset). Notably, the feedforward peak was located at approximately 25 ms delay, longer than the delay of the feedforward peak for the V1-V2 interaction and with a broader profile (compare with Fig. 3b). Over time, the initial feedforward interaction was replaced by a feedback-dominated interaction (Fig. 4b, yellow curve; compare with Fig. 3b; 125 ms after stimulus onset). Figure 4c shows the V1-V4 population correlation functions at all epochs during the trial. The shift from a feedforward- to a feedback-dominated interaction was present for all recording sessions (Fig. 4d; average feedforward ratio, computed in the -50 to 50 ms delay range: 0.088 ± 0.014 SEM for early evoked activity; -0.038 ± 0.008 SEM for late evoked activity; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between early evoked and late evoked activity across all 5 recording sessions; t-test for feedforward ratio, $p = 0.020$ for early evoked activity, $p = 0.003$ for late evoked activity). Importantly, this temporal structure was absent in interactions between subpopulations within each cortical area (Supplementary Fig. 4), and when we shuffled responses to remove the trial-by-trial correspondence between areas (Fig. 4a,b, faded traces).

Population structure of inter-areal interactions

Past work has suggested that inter-areal interactions are selective, in terms of which population activity patterns are related across areas[29,42]. That is, not all activity fluctuations in one area are reflected in the activity of its downstream targets: some fluctuations remain private to the source area. In our analysis thus far, we have focused solely on the strength and directionality of inter-areal interactions.

Given the observed dynamics of inter-areal interactions, we wondered whether the patterns of activity relayed across areas might be different between feedforward- and feedback-dominated periods. One possibility is that the patterns of activity most related across the two areas are similar during these two periods. Since feedback signaling is hypothesized to alter, or correct, visual representations upstream[43–45], one might expect that the dimensions most affected by feedback are the same dimensions that are involved in feedforward interactions. This would suggest feedforward and feedback interactions “read from” and “write to” the same population activity patterns, sharing the same communication channel. Alternatively, feedforward and feedback interactions might unfold through separate channels involving distinct population activity patterns, and thus perhaps minimizing how much they directly interact. This would suggest that feedback processing affects dimensions of upstream activity that are not directly involved in relaying visual information downstream.

To distinguish between these possibilities, we divided the trial in epochs and measured how the canonical dimensions identified during one epoch generalized to another. For example, we asked whether the canonical dimensions identified during the feedforward-dominated period (Fig. 5a) captured inter-areal correlations during the feedback-dominated period as well as the canonical dimensions identified during that feedback-dominated period (Fig. 5b). Good generalization would imply that the same patterns of activity were related across areas during periods of feedforward- and feeback-dominated interactions. If, however, the patterns of activity most related across areas differed, the canonical dimensions found during the feedforward-dominated periods would not capture inter-areal correlations during the feedback-dominated periods (Fig. 5c).

We found that dimensions identified early in the evoked activity period, when V1-V2 interactions were feedforward-dominated, did not generalize well to later epochs (Fig. 6a; average normalized correlation, for which a value of 1 indicates perfect generalization: 0.56 ± 0.05 for mid evoked, 0.59 ± 0.04 for late evoked, 0.36 ± 0.04 for late spontaneous; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between correlation captured using early evoked vs mid evoked, late evoked or late spontaneous dimensions in the corresponding epochs, across all recording sessions). The failure of dimensions identified during the feedforward-dominated period to generalize to the dimensions identified during spontaneous activity suggests that epochs in the feedforward-dominated period involve distinct patterns of population activity compared to epochs in the feedback-dominated period. The generalization was better between epochs later after stimulus onset, when the correlation functions were more symmetric (Fig. 6b; average normalized correlation: 0.64 ± 0.04 for early evoked, 0.94 ± 0.03 for mid evoked, 0.45 ± 0.03 for late spontaneous; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between correlation captured using early evoked vs mid evoked, late evoked or late spontaneous dimensions in the corresponding epochs, across all recording sessions), indicating that the patterns of activity related between areas are stable for mid and late evoked activity. Dimensions identified during epochs of feedback-dominated interaction, during spontaneous activity, failed to generalize to evoked activity (Fig. 6c; average normalized correlation: 0.47 ± 0.06 for early evoked, 0.50 ± 0.03 for mid evoked, 0.52 ± 0.02 for late evoked; paired one-sided Wilcoxon signed-rank test, $p = 0.03$ for difference between correlation captured using early evoked vs mid evoked, late evoked or late spontaneous dimensions in the corresponding epochs, across all recording sessions). These analyses were carefully designed to focus exclusively on changes in the across-area interaction structure, and to be insensitive to changes in the structure of population activity within each area (see Supplementary Fig. 5, Methods, and Supplementary Information).

To gain a more complete picture, we assessed generalization performance between each possible pairing of epochs for defining canonical dimensions (Fig. 6d, vertical axis), and for testing their relevance (horizontal axis). Each row corresponds to a set of canonical dimensions, identified at a particular epoch, and applied to activity at each of the other epochs. The patterns of generalization performance mirror the changes we observed in the temporal profile of the interaction. As the feedforward interaction weakened after stimulus onset (Fig. 3b, compare red and yellow curves), the patterns of activity most related across the two areas changed as well (Fig. 6d, straight arrow, bottom left). Furthermore, the spontaneous activity period, which was more feedback-dominated than the evoked activity period (Fig. 3d), involved different patterns of activity from those involved in the evoked period (Fig. 6d, curved arrow, top right).

We obtained similar results when analyzing V1-V4 activity. V1-V4 interactions transitioned from a feedforward- to a feedback-dominated interaction (Fig. 4), and the dimensions mediating these interactions changed between these epochs as well (Fig. 6e; the number of epochs is smaller here due to the shorter trial duration; one-sided paired Wilcoxon signed-rank test, $p = 0.03$ for difference between correlation captured using late evoked vs early evoked or late spontaneous dimensions in the corresponding epochs, across all recording sessions). Specifically, the V1-V4 interaction became feedback-dominated at the end of the evoked period (Fig. 4d, yellow circles), and this was accompanied by poor generalization between early and late evoked dimensions (Fig. 6e, straight arrow; average normalized correlation for second epoch of evoked activity for V1-V4: 0.27 ± 0.02). In contrast, the V1-V2 interactions shifted more slowly away from a feedforward-dominated interaction after stimulus onset (Fig. 3d, yellow circles). Consistent with this slower transition, V1-V2 dimensions identified soon after stimulus onset generalized better for nearby epochs of evoked activity, compared to V1-V4 (Fig. 6f, straight arrow; averaged normalized correlation for second epoch of evoked activity for V1-V2: 0.62 ± 0.05).

Taken together, our findings suggest feedforward and feedback inter-areal interactions involve different patterns of population activity. In turn, this implies that the aspects of V1 population activity that are relayed downstream are not necessarily the aspects of activity that are most influenced by feedback. Feedforward and feedback processing might thus occur in separate subspaces of population activity, concurrently and through different "channels".

Discussion

We leveraged multi-area recordings to understand the interactions between neuronal population spiking responses in V1 and downstream areas V2 and V4. We found that interactions are feedforward-dominated shortly after stimulus onset, and become feedback-dominated later in the stimulus period and during spontaneous activity. Thus, when a stimulus persists, or when no stimulus is presented, the role of top-down inputs from areas such as V2 and V4 to V1 is more prominent. Furthermore, we found that the population activity patterns most related across areas during feedforward-dominated periods were distinct from those most related during feedback-dominated periods (Fig. 7). This suggests that feedforward and feedback signals involve distinct axes in population activity space, which might allow them to be relayed with minimal direct interference.

In this study, we measured population correlations in activity across areas at different time lags, and we refer to the identified interactions as feedforward or feedback, based on the lags at which population correlations were maximal. The feedforward interactions that we identified from V1 to V2 are likely to reflect direct (i.e., monosynaptic) input for the following reasons. First, our recordings were performed in the output layers of V1 and input layers of V2[19]. Second, the V1-V2 feedforward peak was sharp, and centered at a delay of 2-3 ms (cf. Fig. 3b,e), consistent with the propagation delay between these areas[46,47]. Third, the feedforward peaks identified for the V1-V2 interactions were absent in recording from neuronal populations with poorly aligned receptive fields (cf. Fig. 3e), consistent with specificity of feedforward connections between these areas[37,38]. In contrast, feedback interactions were less temporally precise than the feedforward interactions, suggestive of a longer signaling loop from V2 back to V1 that may involve polysynaptic paths or shared feedback from more distant areas. These feedback interactions were evident both in recordings from populations with aligned or misaligned receptive fields (cf. Fig. 3d), consistent with the broader visuotopic extent of feedback connections[39–41]. For V1-V4 interactions, both the feedforward and feedback interactions were relatively broad (cf. Fig. 4b), which might be explained by the reduced laminar specificity of our recordings in V4 (chronically implanted arrays, compared to movable tetrodes used in V2), and by a larger number of possible paths by which activity can propagate between these two areas[48,49].

In both sets of experiments (V1-V2 in anesthetized animals and V1-V4 recordings in awake animals), we observed that interactions were feedforward-dominated shortly after stimulus onset, but this feedforward component subsided, giving way to feedback-dominated interactions. However, there was one notable difference: V1-V2 interactions became feedback-dominated only after the stimulus offset, whereas V1-V4 became feedback-dominated during the late evoked period. This difference could reflect a stronger influence of feedback signaling in the awake state, a difference in the areas involved (V2 vs. V4), or the layers in which the neuronal populations were recorded. That we saw a feedback-dominated interaction at all in the anesthetized recordings might seem surprising, since activity in higher cortical areas, and therefore top-down inputs, might be expected to be diminished by anesthesia. Although it is unclear whether the feedback-dominated interaction we observed is the same as that in an awake animal, we note that V2 is a major source of feedback to V1[48] and it remains highly responsive under sufentanil anesthesia[12,18,19].

The transition from feedforward- to feedback-dominated interactions during stimulus drive is broadly consistent with inferences drawn from latency measurements. Because V2 depends on input from V1[50], one would expect interactions between the areas to be feedforward-dominated immediately after stimulus onset. Spatial contextual effects in V1, which are thought to arise in part from feedback from higher visual areas[4], are evident 50-100 ms after response onset[11,51,52], consistent with our observation of a shift away from a feedforward-dominated interaction immediately after response onset to a more balanced (V1-V2) or even feedback-dominated (V1-V4) interaction later in the response. While broadly consistent, our observations significantly extend this prior work. In particular, while measurements of onset may provide information about when feedforward and feedback influences begin, they provide little information about their relative influence once both have been engaged. By using population spiking responses, we are able to see network wide changes in the direction of signaling, as a function of stimulus drive.

Our claim that inter-areal interactions switch between being feedforward- or feedback-dominated was not based solely on differences in the time lags at which inter-areal correlations were strongest. It is also supported by our finding that the structure of the population activity that was most correlated between areas was distinct in these different periods. Specifically, we found that the dimensions of population activity that were most related across areas during feedforward signaling periods were distinct from those that were most related during feedback periods. The relevant activity patterns were highly reliable: during spontaneous activity or the sustained epochs of evoked activity, the dimensions of activity that were most correlated across areas were consistent in time. Yet, when networks switched from feedforward to feedback signaling (or vice-versa), the relevant activity patterns changed abruptly.

Determining the population structure of inter-areal interactions requires great care. In particular, it is important to ensure that apparent changes in inter-areal interactions do not arise solely from changes in the structure of activity within each area (see ref. 53, in press). For instance, a change in activity structure within one area might cause the canonical dimensions identified to change, even if the manner in which activity in the two areas is related is unchanged (see Supplementary Information for an extended discussion). To avoid such confounds, we defined interaction structure using across-area covariance, and measured changes in this structure so as to only reflect changes in the activity subspaces in each area spanned by the across-area covariance. In addition, we confirmed that our approach did not detect interaction changes when the across-area covariance was held fixed (Supplementary Fig. 5).

In previous work, we reported that V1 interactions with V2 occur through a communication subspace, which defines which population activity patterns are related across areas[29]. The communication subspace was identified using reduced rank regression (RRR), a dimensionality reduction technique related to CCA but different in its technical details (for a review, see ref. 53, in press). Here we chose to use CCA because it treats the population activity of each area symmetrically. This allows us to study feedforward and feedback influences using the same analysis. In contrast, RRR treats each population differently – one area is labeled the "source" (the independent variable in linear regression) and the other area is labeled the "target" (the dependent variable). Although RRR and CCA need not identify the same dimensions, we found that a communication subspace was also evident when employing CCA. Namely, a smaller number of canonical dimensions was required to capture across area correlations compared to within-area correlations (Supplementary Fig. 6).

How do our observations of feedforward and feedback interactions inform our understanding of how these forms of signaling contribute to cortical function? While the computational role of feedforward signaling has been extensively investigated, the role of feedback is more enigmatic. Feedback signals have been proposed to improve or correct feedforward signals, e.g., by providing prior information about the sensory input[43–45], by providing a prediction of that input (in predictive coding)[54–56], or by signaling deviations from some higher-order "teaching" signal (in biologically plausible backpropagation)[57–59]. We find that inter-areal interactions just after stimulus onset are feedforward. This might be explained by the abrupt transition from one visual environment to another when a stimulus suddenly appears. Assuming the trial structure is not learned by the visual cortex, stimulus onset is unpredicted or unexpected; according to predictive coding principles, such input should give rise to potent feedforward signaling. As the stimulus persists, inter-areal interactions become feedback-dominated. This transition might indicate that higher cortex is providing signals that attempt to 'explain away' the constant, persistent visual input, and thereby reduce responsivity in lower cortex. Interactions are also feedback-dominated during spontaneous activity. This finding is consistent with proposals that sensory representations combine prior information from higher cortex with sensory drive from the periphery. In the absence of overt visual input (i.e., during spontaneous activity), one would expect responses to reflect more strongly the prior, which would be evident as a top-down dominant interaction.

Our finding that feedforward and feedback interactions involve different patterns of population activity may offer a solution to a central enigma in proposals of how feedback contributes to sensory processing: feedback that is too weak may fail to properly modify representations of the sensory stimulus, but feedback that is too strong may contaminate the representation and lead to hallucinations. One solution for providing robust feedback but allowing some flexibility in how it interacts with the bottom-up sensory representation could be to have these occupy different dimensions of V1 population activity, as we find. The presence of the feedback signal in a target area can then be decoupled from the strength of its influence. This would suggest that the balance between feedforward and feedback signaling in sensory cortex might be achieved using the same principles used by motor cortex to generate preparatory signals without causing muscle contractions[42], by prefrontal networks that host competing sensory inputs but can flexibly switch which one drives the local activity[60], or by visual cortical areas to selectively communicate[29].

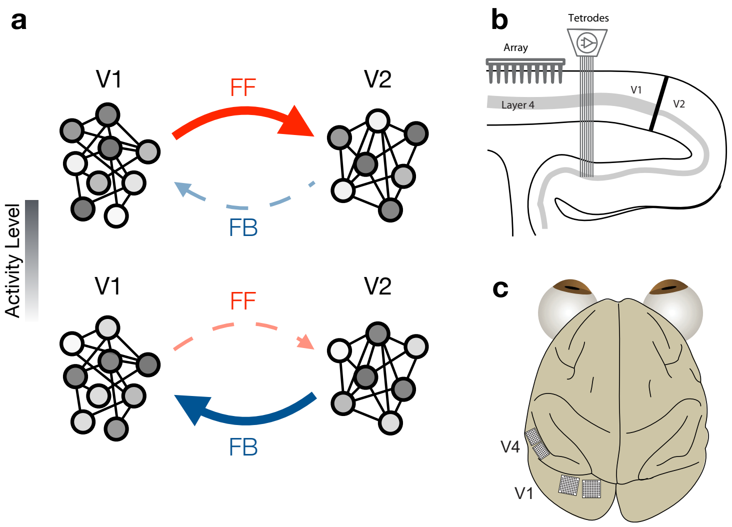


Figure 1 Studying feedforward and feedback interactions using neuronal population activity. **(a)** Each circle represents a neuron in each area, with the shading representing the activity level of the neuron. The population activity patterns involved in feedforward signaling (top) might be distinct from those involved in feedback interactions (bottom). **(b)** Schematic showing a sagittal section of occipital cortex and the recording setup for the V1-V2 recordings. We simultaneously recorded V1 population activity using a 96-channel array and V2 population activity using a set of movable electrodes and tetrodes. **(c)** Schematic showing an overhead view of the recording setup for the V1-V4 awake recordings. We simultaneously recorded V1 and V4 population activity using one 96-channel and one 48-channel array in V1 and a 48-channel array in V4 in the first animal, and two 96-channel arrays in V1 and two 48-channel array in V4 in the second animal.

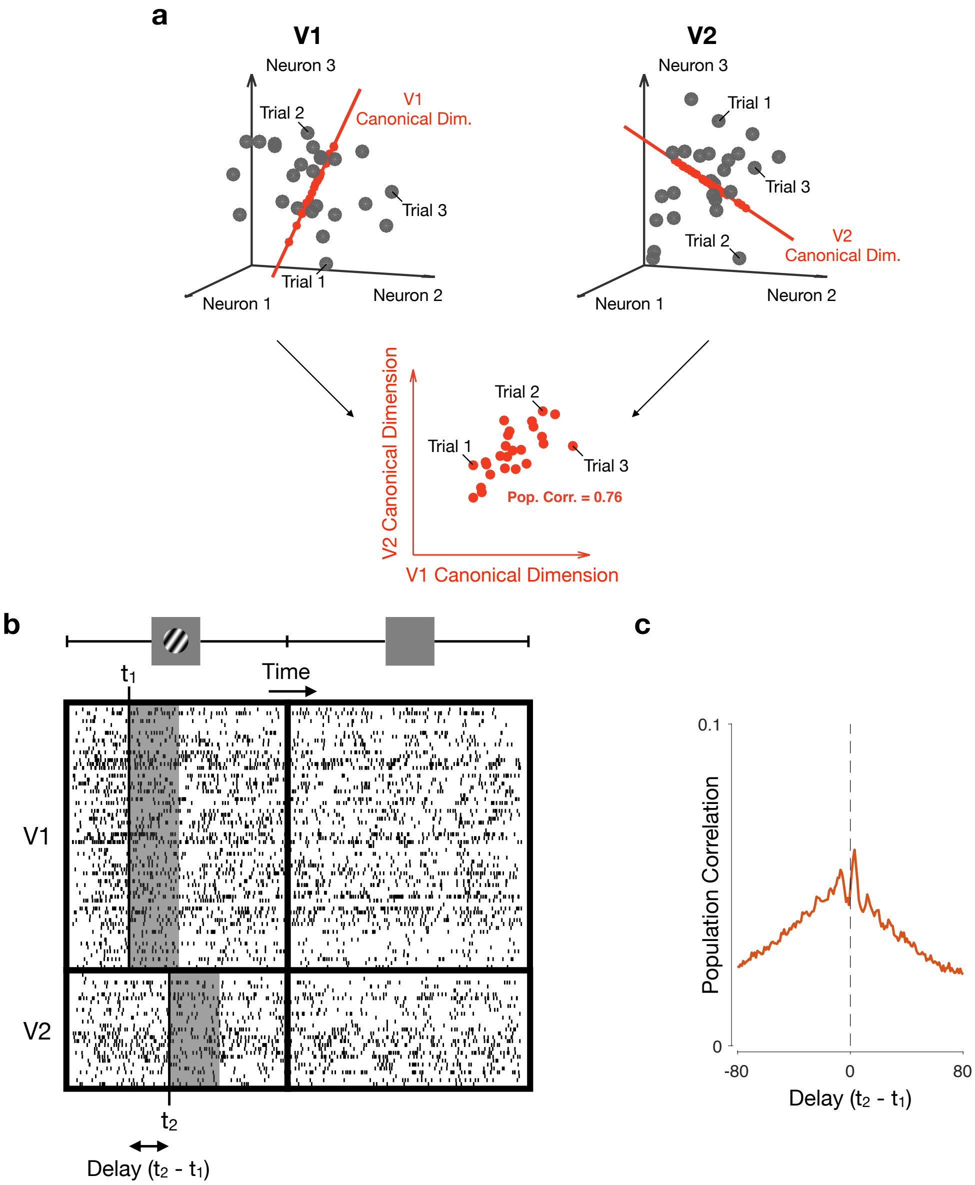


Figure 2 Using Canonical Correlation Analysis (CCA) to capture population interactions. **(a)** Relating activity across two neuronal populations. Each circle represents the population activity recorded on a given trial. For each activity point observed in the V1 population (left panel; gray dots), there is a corresponding, simultaneously recorded activity point observed in V2 (right panel, gray dots). The red axes represent the first pair of canonical dimensions, identified using CCA. Neuronal activity projected onto the first pair of canonical dimensions (red dots) is highly correlated across the two areas (bottom panel). **(b)** Spike counts across the recorded neurons are taken in specified time windows

(gray boxes), which may either be positioned at the same time in both areas (i.e., $t_1 = t_2$) or with a delay between areas ($t_1 \neq t_2$). The activity in each gray box is represented by a circle in panel (a). **(c)** The population correlation function corresponds to the correlation between areas returned by CCA (the correlation associated with the first pair of canonical dimensions), as a function of the time delay between areas ($t_2 - t_1$).

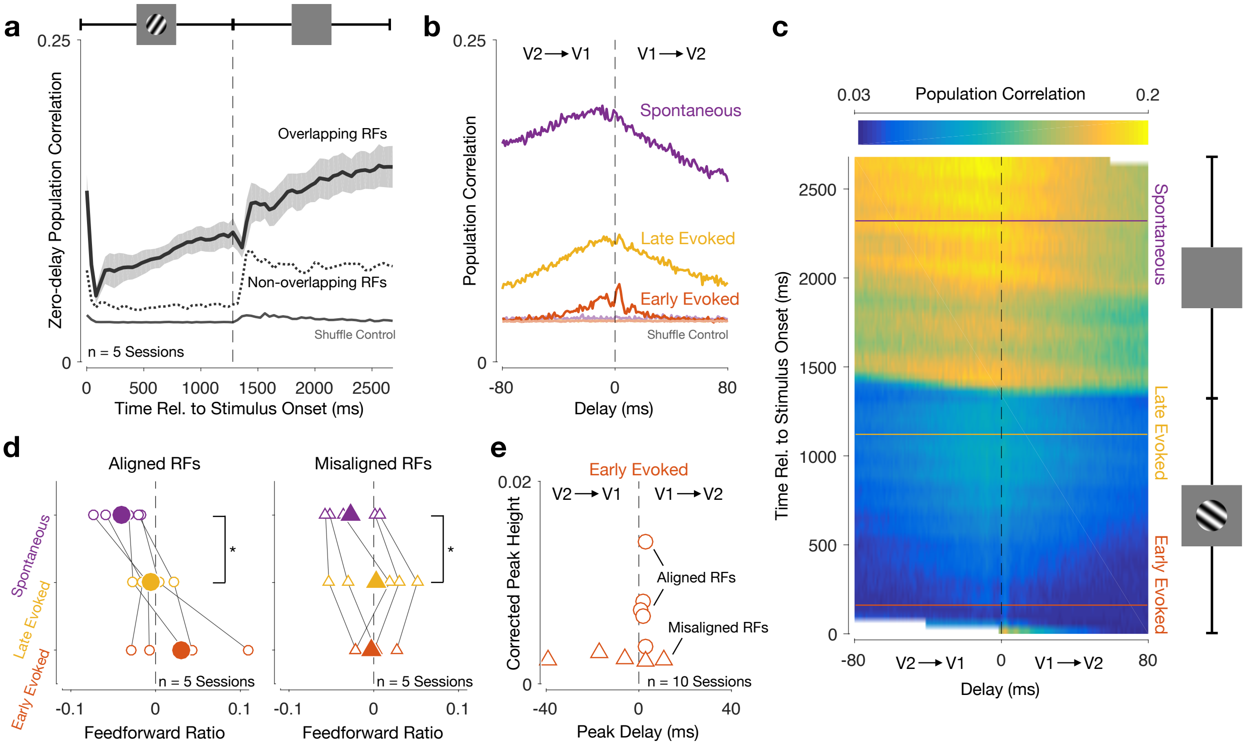


Figure 3 V1-V2 interaction transitions from feedforward-dominated shortly after stimulus onset to feedback-dominated during the spontaneous period. **(a)** Inter-areal zero-delay population correlation increased throughout the trial, and was higher for spontaneous activity than for evoked activity. Zero-delay refers to spike counts taken in the same time window in the two areas ($t_1 = t_2$ in Fig. 2b). Black line shows the average across all recording sessions for which the V1 and V2 populations have aligned receptive fields. Shading indicates S.E.M. Dotted line shows average across all recording sessions where the the V1 and V2 receptive fields are misaligned. Gray line shows average population correlation after shuffling trial correspondence between the two areas. **(b)** Population correlation functions for an example session (red: early evoked, yellow: late evoked; purple: spontaneous). Faded lines show population correlation functions after shuffling trial correspondence between the two areas (note that there are multiple superimposed lines). **(c)** Population correlations at all times during the trial. The horizontal axis represents the time delay between areas ($t_1 - t_2$), and the vertical axis

represents time relative to stimulus onset (t_1). Horizontal lines (red, yellow, and purple) indicate epochs used in panel (b). Dashed vertical line indicates zero-delay population correlations shown in panel (a). White area denotes times for which population correlations could not be computed: the V2 activity window had reached either the beginning or the end of the trial. Same session as in panel (b). **(d)** Feedforward ratio for different epochs of evoked and spontaneous activity. Left panel shows sessions for which the V1 and V2 populations have aligned receptive fields; right panel shows sessions where the the V1 and V2 receptive fields are misaligned. Solid symbols show the average across all recording sessions, whereas open symbols correspond to each recording session. **(e)** An early feedforward peak is only present in recording sessions where the V1 and V2 populations have aligned receptive fields. Peak height is measured after performing a jitter-correction to isolate fast timescale interactions (see Methods). Circles correspond to recording sessions for which the V1 and V2 populations have aligned receptive fields. Triangles correspond to sessions in which the V1 and V2 receptive fields are misaligned.

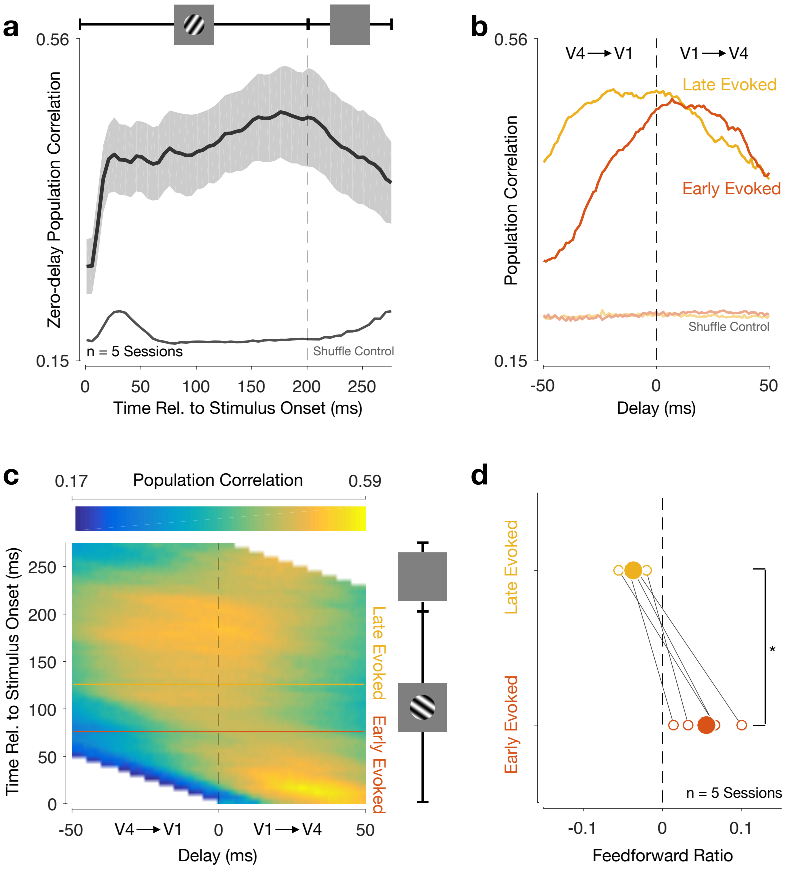


Figure 4 V1-V4 interaction transitions from feedforward- to feedback-dominated during the evoked period. **(a)** Inter-areal zero-delay population correlation increased throughout the evoked period. Black line shows average across all recording sessions. Shading indicates S.E.M. Gray line shows average population correlations after shuffling trial correspondence between the two areas. **(b)** Population correlation functions for an example session, for early (red) and late evoked (yellow) activity. Due to the short duration of the inter-stimulus period, we could not compute a population correlation function for spontaneous activity. Faded lines show population correlation functions after shuffling trial correspondence between the two areas (note that there are multiple superimposed lines). **(c)** Population correlations at all times during the trial. The horizontal axis represents the time delay between areas ($t_1 - t_2$), and the vertical axis represents time relative to stimulus onset (t_1). Horizontal lines (red and yellow) indicate epochs used in panel (b). Dashed vertical line indicates zero-delay population correlations

shown in panel (a). White area denotes times for which population correlations could not be computed: the V4 activity window had reached either the beginning or the end of the trial. Same session as in panel (b). **(d)** Feedforward ratio for early and late evoked activity. Solid circles show the average across all recording sessions, whereas open circles correspond to each recording session.

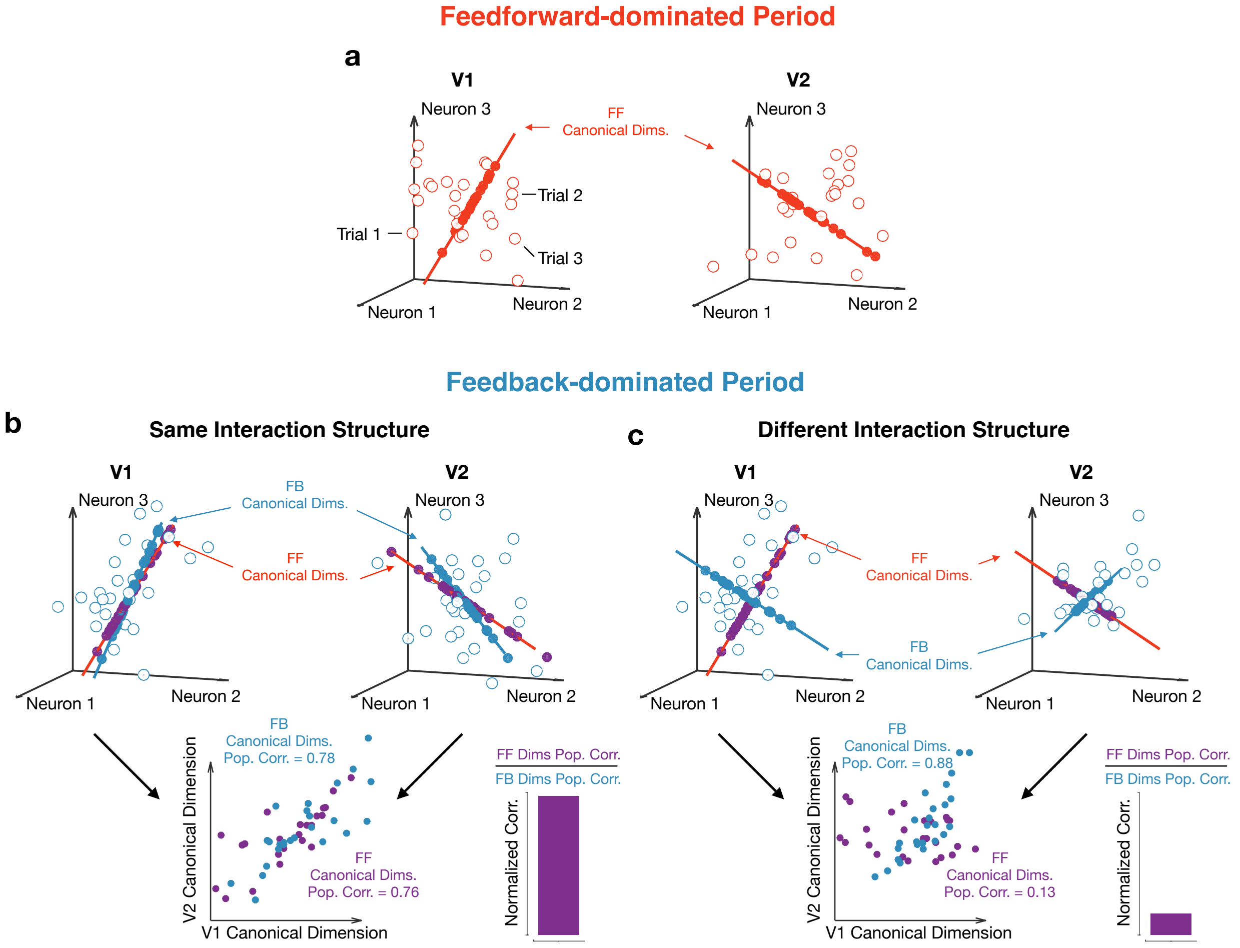


Figure 5 Assessing whether feedforward- and feedback-dominated interactions involve the same population activity patterns. **(a)** Canonical dimensions identified during a feedforward-dominated period in the trial (red dimensions). These are putative "Feedforward" (FF) canonical dimensions. Open red circles denote activity during the feedforward-dominated period. Solid red circles denote the projection onto the FF canonical dimensions. **(b)** We can then ask whether these FF canonical dimensions generalize to a feedback-dominated period. One possibility is that the interaction structure (defined using the canonical dimensions) remains stable across the two periods. In this case, the FF canonical dimensions (red dimensions) capture a similar level of correlation during the feedback-dominated period as the canonical dimensions identified during this period, the putative "Feedback" (FB) canonical dimensions (blue dimensions). As a result,

the normalized correlation, the ratio of the population correlation for the FF canonical dimensions to that for the FB canonical dimensions (both computed in a cross validated manner; see Methods), is close to 1. Open blue circles denote activity during the feedback-dominated period. Solid purple circles denote the projection of activity during the feedback-dominated period onto the FF canonical dimensions. Solid blue circles denote the projection onto the FB canonical dimensions. **(c)** Alternatively, the interaction structure might change across the two periods. In this case, the FF dimensions capture only a small fraction of the population correlation during the feedback-dominated period. Same conventions as in panel (b).

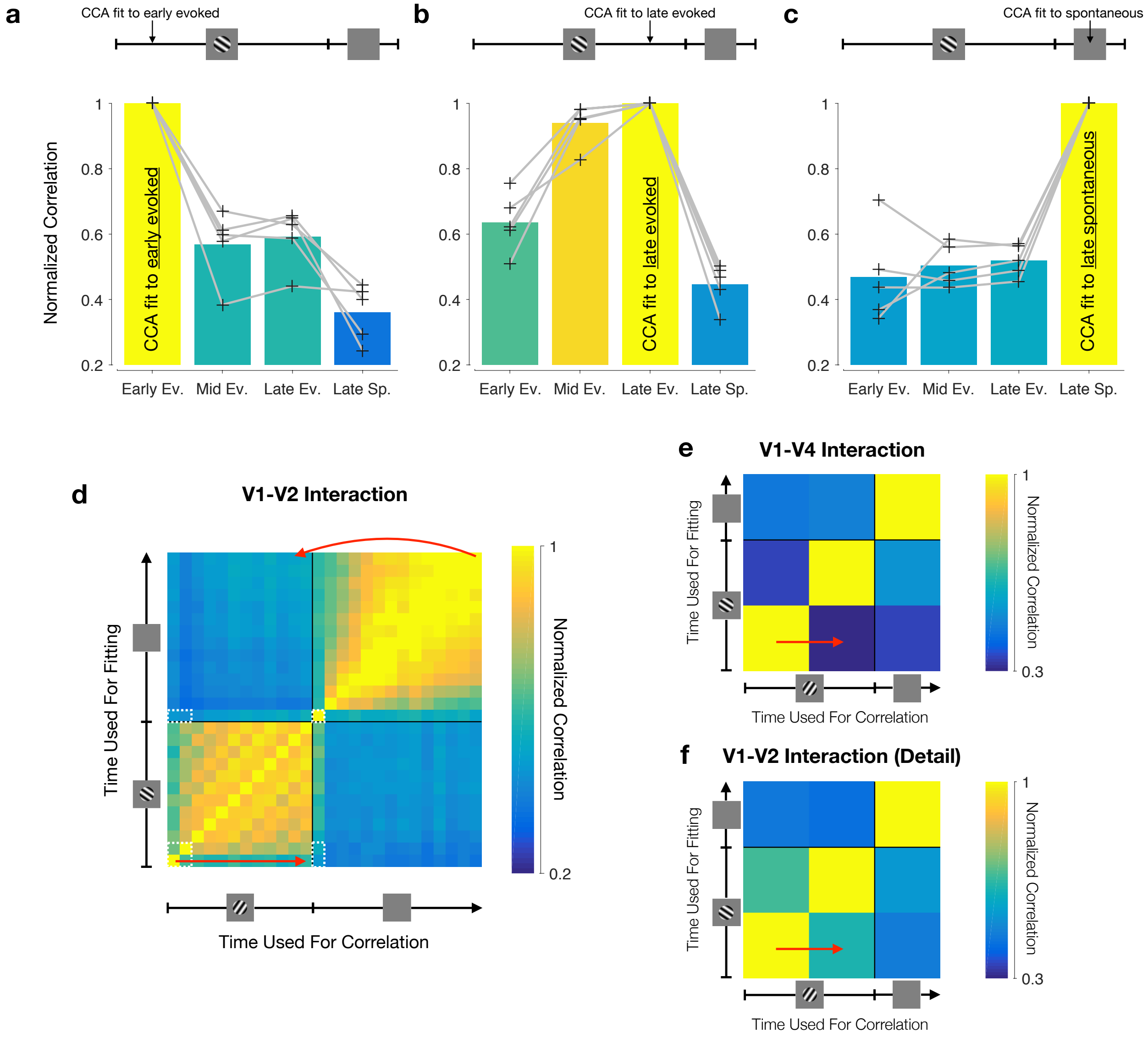


Figure 6 Interaction structure is distinct for the feedforward- and feedback-dominated periods. **(a)** The dimensions found by fitting CCA shortly after stimulus onset (80 ms after stimulus onset) do not generalize well to later epochs in the evoked period, and worse still during the spontaneous period. Grey lines correspond to each of the 5 recording sessions. We report the normalized correlation, defined as the total correlation captured at the test epoch by the dimensions fit to some other epoch over the total correlation captured by the dimensions fit to the test epoch (both computed in a cross-validated manner; see Methods). **(b)** Dimensions identified late in the evoked period (1180 ms after stimulus onset) do not

generalize well to early evoked epochs and to epochs in the spontaneous period, but generalize well to mid-evoked activity. Same conventions as in panel (a). **(c)** Dimensions identified during the spontaneous period do not generalize well to the evoked period. Same conventions as in panel (a). **(d)** Assessing changes in interaction structure across the entire trial. The trial was divided into 100 ms segments, and CCA was applied separately to the activity in each time window. The top two canonical pairs associated with each window were then used to capture inter-areal correlations in the other time windows (see Methods). Each row corresponds to the time during the trial during which the canonical dimensions were identified. Each column corresponds to the time during the trial where the population correlation is assessed. Each entry shows the average across all recording sessions. Straight arrow highlights the comparison of the interaction structure within the evoked period. Curved arrow highlights the comparison of the interactions structure between the spontaneous and the evoked periods. Dashed white boxes indicate epochs reproduced in panel (f). **(e)** Comparing identified dimensions across epochs for the awake V1-V4 recordings. The trial was divided into 100 ms segments, and CCA was applied separately to the activity in each time window. The top canonical pair associated with each window was then used to capture inter-areal correlations in the other time windows (see Methods). Arrow highlights the comparison of the interaction structure within the evoked period. Same conventions as in panel (d). **(f)** Detailed view of the V1-V2 generalization performance for the comparable epochs between the V1-V2 and V1-V4 recordings. Epochs are indicated by the dashed white boxes in panel (d).

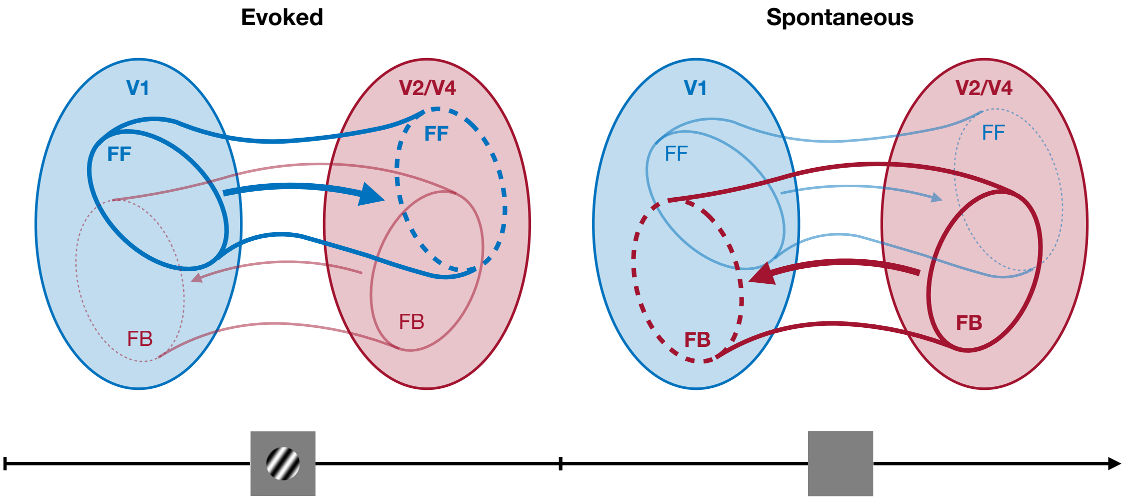


Figure 7 Summary of results. During the early evoked period, interactions between areas tend to be feedforward-dominated. Later during the evoked period and during the spontaneous period, interactions between areas become feedback-dominated. Furthermore, feedforward- and feedback-dominated interactions involve different population activity patterns. Larger ellipses represent the set of all activity patterns one might observe in either the V1 or the V2/V4 populations. The smaller ellipses represent the activity patterns most related across the two areas.

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. Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810–824.e9 (2019).

3. Lamme, V. A., Supr, H. & Spekreijse, H. Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology* **8**, 529–535 (1998).

4. Angelucci, A. & Bressloff, P. C. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. In Martinez-Conde (ed.) *Progress in Brain Research*, vol. 154, Part A of *Visual PerceptionFundamentals of Vision: Low and Mid-Level Processes in Perception*, 93–120 (Elsevier, 2006).

5. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nature Reviews Neuroscience* **14**, 350–363 (2013).

6. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503**, 51–58 (2013).

7. Schmolesky, M. T. *et al.* Signal Timing Across the Macaque Visual System. *Journal of Neurophysiology* **79**, 3272–3278 (1998).

8. de Lafuente, V. & Romo, R. Neural correlate of subjective sensory experience gradually builds up across cortical areas. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14266–14271 (2006).

9. Herndez, A. *et al.* Decoding a Perceptual Decision Process across Cortex. *Neuron* **66**, 300–314 (2010).

10. Siegel, M., Buschman, T. J. & Miller, E. K. Cortical information flow during flexible sensorimotor decisions. *Science* **348**, 1352–1355 (2015).

11. Supèr, H., Spekreijse, H. & Lamme, V. A. F. Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature Neuroscience* **4**, 304–310 (2001).

12. Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience* **16**, 974–981 (2013).

13. Chen, M. *et al.* Incremental Integration of Global Contours through Interplay between Visual Cortical Areas. *Neuron* **82**, 682–694 (2014).

14. Schwiedrzik, C. M. & Freiwald, W. A. High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron* **96**, 89–97.e4 (2017).

15. Issa, E. B., Cadieu, C. F. & DiCarlo, J. J. Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *eLife* **7**, e42870 (2018).

16. Roe, A. W. & Ts'o, D. Y. Specificity of Color Connectivity Between Primate V1 and V2. *Journal of Neurophysiology* **82**, 2719–2730 (1999).

17. Nowak, L. G., Munk, M. H. J., James, A. C., Girard, P. & Bullier, J. Cross-Correlation Study of the Temporal Interactions Between Areas V1 and V2 of the Macaque Monkey. *Journal of Neurophysiology* **81**, 1057–1074 (1999).

18. Jia, X., Tanabe, S. & Kohn, A. Gamma and the Coordination of Spiking Activity in Early Visual Cortex. *Neuron* **77**, 762–774 (2013).

19. Zandvakili, A. & Kohn, A. Coordinated Neuronal Activity Enhances Corticocortical Communication. *Neuron* **87**, 827–839 (2015).

20. Campo, A. T. *et al.* Task-driven intra- and interarea communications in primate cerebral cortex. *Proceedings of the National Academy of Sciences* **112**, 4761–4766 (2015).

21. Campo, A. T. *et al.* Feed-forward information and zero-lag synchronization in the sensory thalamocortical circuit are modulated during stimulus perception. *Proceedings of the National Academy of Sciences* **116**, 7513–7522 (2019).

22. Gregoriou, G. G., Gotts, S. J., Zhou, H. & Desimone, R. High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention. *Science* **324**, 1207–1210 (2009).

23. Salazar, R. F., Dotson, N. M., Bressler, S. L. & Gray, C. M. Content-Specific Fronto-Parietal Synchronization During Visual Working Memory. *Science* **338**, 1097–1100 (2012).

24. van Kerkoerle, T. *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 14332–14341 (2014).

25. Bastos, A. M., Vezoli, J. & Fries, P. Communication through coherence with inter-areal delays. *Current Opinion in Neurobiology* **31**, 173–180 (2015).

26. Truccolo, W., Hochberg, L. R. & Donoghue, J. P. Collective dynamics in human and monkey sensorimotor cortex: predicting single neuron spikes. *Nature Neuroscience* **13**, 105–111 (2010).

27. Chen, J. L., Voigt, F. F., Javadzadeh, M., Krueppel, R. & Helmchen, F. Long-range population dynamics of anatomically defined neocortical networks. *eLife* **5**, e14679 (2016).

28. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).

29. 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).

30. Ruff, D. A. & Cohen, M. R. Simultaneous multi-area recordings suggest that attention improves performance by reshaping stimulus representations. *Nature Neuroscience* **22**, 1669–1676 (2019).

31. Perich, M. G., Gallego, J. A. & Miller, L. E. A Neural Population Mechanism for Rapid Learning. *Neuron* **100**, 964–976.e7 (2018).

32. Ames, K. C. & Churchland, M. M. Motor cortex signals for each arm are mixed across hemispheres and neurons yet partitioned within the population response. *eLife* **8** (2019).

33. Veuthey, T. L., Derosier, K., Kondapavulur, S. & Ganguly, K. Single-trial cross-area neural population dynamics during long-term skill learning. *Nature Communications* **11**, 4057 (2020).

34. Kohn, A. *et al.* Principles of Corticocortical Communication: Proposed Schemes and Design Considerations. *Trends in Neurosciences* (2020).

35. Harrison, M. T. & Geman, S. A Rate and History-Preserving Resampling Algorithm for Neural Spike Trains. *Neural Computation* **21**, 1244–1258 (2008).

36. Smith, M. A. & Kohn, A. Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. *The Journal of Neuroscience* **28**, 12591–12603 (2008).

37. Rockland, K. S. & Pandya, D. N. Cortical connections of the occipital lobe in the rhesus monkey: Interconnections between areas 17, 18, 19 and the superior temporal sulcus. *Brain Research* **212**, 249–270 (1981).

38. Salin, P. A. & Bullier, J. Corticocortical connections in the visual system: structure and function. *Physiological Reviews* **75**, 107–154 (1995).

39. Rockland, K. S. & Virga, A. Terminal arbors of individual Feedback axons projecting from area V2 to V1 in the macaque monkey: A study using immunohistochemistry of anterogradely transported Phaseolus vulgaris-leucoagglutinin. *Journal of Comparative Neurology* **285**, 54–72 (1989).

40. Angelucci, A. *et al.* Circuits for Local and Global Signal Integration in Primary Visual Cortex. *Journal of Neuroscience* **22**, 8633–8646 (2002).

41. Shmuel, A. *et al.* Retinotopic Axis Specificity and Selective Clustering of Feedback Projections from V2 to V1 in the Owl Monkey. *Journal of Neuroscience* **25**, 2117–2131 (2005).

42. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* **17**, 440–448 (2014).

43. Haefner, R., Berkes, P. & Fiser, J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* **90**, 649–660 (2016).

44. Orban, G., Berkes, P., Fiser, J. & Lengyel, M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron* **92**, 530–543 (2016).

45. Aitchison, L. & Lengyel, M. With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology* **46**, 219–227 (2017).

46. Girard, P., Hup, J. M. & Bullier, J. Feedforward and Feedback Connections Between Areas V1 and V2 of the Monkey Have Similar Rapid Conduction Velocities. *Journal of Neurophysiology* **85**, 1328–1331 (2001).

47. El-Shamayleh, Y., Kumbhani, R. D., Dhruv, N. T. & Movshon, J. A. Visual Response Properties of V1 Neurons Projecting to V2 in Macaque. *Journal of Neuroscience* **33**, 16594–16605 (2013).

48. Felleman, D. J. & Essen, D. C. V. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex* **1**, 1–47 (1991).

49. Markov, N. T. *et al.* Cortical High-Density Counterstream Architectures. *Science* **342** (2013).

50. Girard, P. & Bullier, J. Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *Journal of Neurophysiology* **62**, 1287–1302 (1989).

51. Bair, W., Cavanaugh, J. R. & Movshon, J. A. Time Course and Time-Distance Relationships for Surround Suppression in Macaque V1 Neurons. *Journal of Neuroscience* **23**, 7690–7701 (2003).

52. Smith, M. A., Bair, W. & Movshon, J. A. Dynamics of Suppression in Macaque Primary Visual Cortex. *Journal of Neuroscience* **26**, 4826–4834 (2006).

53. Semedo, J. D., Gokcen, E., Machens, C. K., Kohn, A. & Yu, B. M. Statistical methods for dissecting interactions between brain areas. *Current Opinion in Neurobiology* (2020).

54. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**, 79–87 (1999).

55. Friston, K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 815–836 (2005).

56. Keller, G. B. & Mrsic-Flogel, T. D. Predictive Processing: A Canonical Cortical Computation. *Neuron* **100**, 424–435 (2018).

57. Sacramento, J., Ponte Costa, R., Bengio, Y. & Senn, W. Dendritic cortical microcircuits approximate the backpropagation algorithm. In Bengio, S. *et al.* (eds.) *Advances in Neural Information Processing Systems 31*, 8721–8732 (Curran Associates, Inc., 2018).

58. Whittington, J. C. R. & Bogacz, R. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences* **23**, 235–250 (2019).

59. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nature Reviews Neuroscience* **21**, 335–346 (2020).

60. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

Acknowledgements We thank E. Gokcen and A. Motiwala for invaluable discussions and for providing feedback on the manuscript. This work was supported by the Fundação para a Ciência e a Tecnologia graduate scholarship SFRH/BD/52069/2012 (J.D.S.), John and Claire Bertucci Graduate Fellowship (J.D.S.), NIH U01 NS094288 (C.K.M.), Simons Collaboration on the Global Brain 364994 (B.M.Y., A.K.), 543009 (C.K.M.), 543065 (B.M.Y.), 542999 (A.K.), NIH R01 HD071686 (B.M.Y.), NIH CRCNS R01 NS105318 (B.M.Y.), NIH CRCNS R01 MH118929 (B.M.Y.), NIH R01 EB026953 (B.M.Y.), NSF NCS BCS 1533672 and 1734916 (B.M.Y.), NIH EY016774 (A.K.).

Author Contributions J.D.S., C.K.M., A.K. and B.M.Y. designed the analyses. J.D.S. performed all the analyses. A.I.J., A.Z., A.A. and A.K. designed and performed the experiments. J.D.S., C.K.M., A.K. and B.M.Y. wrote the manuscript. C.K.M., A.K. and B.M.Y. contributed equally to this work.

Methods

Recordings and visual stimulation

Anesthetized V1-V2

Animal procedures and recording details have been described in previous work[19,36]. Briefly, animals (macaca fascicularis, male, 2-3 years old) were anesthetized with ketamine (10 mg/kg) and maintained on isoflurane (1%-2%) during surgery. Recordings were performed under sufentanil (typically 6-18 mg/kg/hr) anesthesia. Vecuronium bromide (150 mg/kg/hr) was used to prevent eye movements. The duration of each experiment (which comprised multiple recording sessions) varied from 5 to 7 days. All procedures were approved by the IACUC of the Albert Einstein College of Medicine.

The data analyzed here are those reported in ref. 29, and a subset of recording sessions reported in ref. 19. Activity in V1 was recorded using a 96 channel Utah array (400 micron inter-electrode spacing, 1 mm length, inserted to a nominal depth of 600 microns; Blackrock, UT). We recorded V2 activity using a set of electrodes/tetrodes (interelectrode spacing 300 microns) whose depth could be controlled independently (Thomas Recording, Germany). These electrodes were lowered through V1, the underlying white matter, and then into V2. Within V2, we targeted neurons in the input layers. We verified the recordings were performed in the input layers using measurements of the depth in V2 cortex, histological confirmation (in a subset of recordings), and correlation measurements. For complete details see ref. 19. Voltage snippets that exceeded a user-defined threshold were digitized and sorted offline. The sampled neurons had spatial receptive fields within 2-4 deg of the fovea, in the lower visual field.

We measured responses evoked by drifting sinusoidal gratings (1 cyc/deg ; drift rate of 3-6.25 Hz; 2.6-4.9 deg in diameter; full contrast, defined as Michelson contrast, $(L_{max} - L_{min})/(L_{max} + L_{min})$, where L_{min} is 0 cd/m^2 and L_{max} is 80 cd/m^2) at 8 different orientations (22.5 deg steps), on a calibrated CRT monitor placed 110 cm from the animal (1024 x 768 pixel resolution at a 100 Hz refresh rate; EXPO). Each stimulus was presented 400 times for 1.28 s. Each presentation was followed by an interval of 1.5 s during which a gray screen was presented.

We recorded neuronal activity in three animals. In two of the animals, we recorded in two different but nearby locations in V2, providing distinct middle-layer populations, yielding a total of five recording sessions.

Awake V1-V4

Animal procedures and methods have been reported previously in previous work[61]. In brief, animals (two male, adult cynomolgus macaques) were trained to maintain fixation on a small spot (0.2 x 0.2 deg, 80 cd/m2) on a gray background (40 cd/m2) within a 1.08-1.4 degree diameter fixation window. Eye-position was monitored using a video tracking system (Eyelink II, SR research, ON, Canada) with a sampling rate of 500 Hz. Stimuli were presented on a calibrated monitor 64 cm away from the animal (1024 x 768 resolution for monkey 1, 1400x1050 for monkey 2; 100 Hz refresh rate). After training, Utah arrays (0.4 mm spacing; 1 mm electrode length, Blackrock, UT) were implanted in V1 and V4. For monkey 1 we implanted one 96 channel and one 48 channel array in V1 and one 48 channel array in V4. Monkey 2 had two 96 channel arrays in V1 and two 48 channel arrays in V4 (see Fig. 1c). We targeted the arrays to have matching retinotopic locations in V1 and V4 by relying on anatomical markers and previous mapping studies. Receptive fields were in the lower right visual hemifield and largely overlapping for V1 and V4 populations in both monkeys (Supplementary Fig. 1). All procedures were approved by the IACUC of the Albert Einstein College of Medicine.

Extracellular voltage signals were amplified and band-pass filtered between 250 and 7.5 kHz using commercial acquisition software (Blackrock Microsystems, UT and Grapevine, Ripple, UT). Voltage snippets that exceeded a user-defined threshold were digitized and sorted offline.

Visual stimuli and task contingencies were presented using custom openGL software (EXPO). We used full-contrast sinusoidal drifting gratings (spatial frequency 2 cyc/deg; drift rate: 5 Hz). Stimulus position and diameter were chosen to maximize visual responses. Stimulus diameter was set to 2.5 deg for monkey 1 and 7 deg for monkey 2. Each recording session involved four grating orientations, chosen such that there were two pairs of orientations 5 deg apart, and 90 deg between the two pairs (e.g., 0, 5, 90, 95 deg).

Trials began with the animal fixating on a small spot in the center of the screen. After a delay of 100 ms we presented a random series of gratings (three for monkey 1, four for monkey 2). Each stimulus presentation lasted for 200 ms and was followed by an inter-stimulus interval of 150 ms (grey screen). Animals were positively reinforced with a liquid reward if fixation was maintained throughout the trial. Animals performed on average 1080 ± 255 trials, resulting in 3721 ± 1081 stimulus presentations per session. We recorded neural activity for three sessions in monkey 1 and two sessions in monkey 2.

Data preprocessing

Anesthetized V1-V2

In order to capture how moment-to-moment fluctuations in spiking activity were related across the two areas, we subtracted the corresponding peri-stimulus time histogram (PSTH) from each spike train, which was computed separately for each neuron and grating orientation (after z-scoring the activity of each neuron separately for each of the 8 grating orientations). The PSTH was computed across the entire trial period, including the stimulus presentation period and the subsequent inter-trial period. The resulting residual activity was then pooled across all 8 grating orientations for each recording session. These residual fluctuations can be interpreted as perturbations of the "signal", or mean activity across trials. By focusing on perturbations of the signal, we can then use linear methods such as CCA (see below) as a local linear approximation to what is likely a globally non-linear relationship of activity across areas[29,62]. For all analyses, we excluded neurons that fired less than 0.5 spikes/s on average across all trials.

Awake V1-V4

To minimize the influence of adaptation effects, we analyzed activity across only the second and third grating presentations, for which V1-V4 responses were qualitatively similar (and smaller than the response to the first stimulus presentation). Activity for each neuron was z-scored separately for the second and third grating presentations, and for each of the 4 grating orientations. As with the V1-V2 recordings, we subtracted the corresponding PSTH from each trial, which was computed separately for each neuron and stimulus condition (i.e., combination of grating orientations). The PSTH was computed across the entire trial period, including the stimulus presentation period and the subsequent inter-trial period. The resulting residual activity was then pooled across all stimulus conditions for each recording session. We observed cross-talk between a small proportion of electrode pairs (average across recording sessions: $1.3\% \pm 0.8\%$ SEM), evident as a surfeit (> 0.025 coincidences/spike) of precise (0.1 ms) synchronous events.

We addressed this by removing one of the electrodes in each affected pair. For all analyses, we excluded neurons that fired less than 0.5 spikes/s on average, across all trials.

Population correlation functions

When computing the population correlation functions for the anesthetized V1-V2 recordings (Fig. 3), we sought to focus on fast time-scale interaction effects. For this reason, we counted spikes in 1 ms non-overlapping bins. For the awake V1-V4 recordings (Fig. 4), due to the smaller number of trials per recording session and the longer conduction delay between V1 and V4[7] we counted spikes in non-overlapping 25 ms bins.

Interaction structure analysis

For the interaction structure analysis (Fig. 6), for which we were interested in estimating the activity patterns most correlated across areas, we counted spikes in 100 ms non-overlapping bins. Activity was binned starting 50 ms after stimulus onset and extending until the end of the stimulus presentation period (1.2 s of evoked activity) and then starting 50 ms after stimulus offset and extending until the end of the inter-trial period (1.45s of spontaneous activity). We used larger time bins than for computing population correlation functions to increase the reliability of the estimated population activity patterns, in exchange for less temporal resolution. Likewise, for the awake V1-V4 recordings we counted spikes in 100 ms non-overlapping bins, starting 50 ms after stimulus onset and extending until the end of the trial (150 ms of evoked activity and 150 ms of spontaneous activity, for a total of 300 ms).

Population correlation analysis

In order to capture population correlations between cortical areas, we used Canonical Correlation Analysis (CCA)[63]. CCA finds pairs of dimensions, one in each area, such that the correlation between the projected activity onto these dimensions is maximally correlated:

$$\arg\max_{\mathbf{a},\mathbf{b}} \operatorname{corr}(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b})$$

where $\mathbf{X}$ is a $n \times p_x$ matrix containing the residual activity in the V1 population, $\mathbf{Y}$ is a $n \times p_y$ matrix containing the residual activity in the V2 (or V4) population, n represents the number of data points, and p_x and p_y are the number of recorded neurons in each of the two areas, respectively. The vectors $\mathbf{a}$ and $\mathbf{b}$ have dimensions $p_x \times 1$ and $p_y \times 1$, respectively defining dimensions in the population activity space of each area. CCA can find additional pairs of dimensions, by requiring that subsequent pairs are uncorrelated with those previously identified.

In order to measure population correlations at different epochs in the trial, and at different time delays between the areas, we defined two windows of activity, one in each area. Window length was 80 ms for the V1-V2 recordings, and 75 ms for the V1-V4 recordings. Activity was then binned inside each window using 1 ms bins for the V1-V2 recordings (80 data points per window), and 25 ms for the V1-V4 recordings (3 data points per window). The reported results were robust to the specific binning and window length chosen, over a reasonable range.

CCA was then applied to the residual activity taken from all trials within these windows. Given two windows of activity starting at times t_1 and t_2 (relative to the start of the trial), $\mathbf{X}_{t_1}$ and $\mathbf{Y}_{t_2}$, the population correlation between the two areas is given by:

$$P(t_1, t_2) = \max_{\mathbf{a},\mathbf{b}} \text{corr}(\mathbf{X}_{t_1}\mathbf{a}, \mathbf{Y}_{t_2}\mathbf{b})$$

Defining the time within the trial as $t = t_1$ and the delay between the activity in the two areas as $d = t_2 - t_1$, each entry in the population correlation function is given by:

$$C(t, d) = P(t, t + d) = \max_{\mathbf{a},\mathbf{b}} \text{corr}(\mathbf{X}_t\mathbf{a}, \mathbf{Y}_{t+d}\mathbf{b})$$

CCA tended to identify only one pair of dimensions with highly significant population correlations: correlations associated with the second canonical pair were on average 60% lower than for the first pair and close to chance level. As such, we constructed the population correlation functions using the first pair of canonical dimensions.

To isolate fast-timescale features in the early evoked activity (Fig. 3e), we computed jitter-corrected population correlation functions. To do so, we jittered the spike times (25 ms jitter window) following the procedure in ref. 36. We then computed population correlation functions using 1 ms binning and a window length of 480 ms, starting 80 ms after stimulus onset, for both the residual activity and the jittered activity. Finally, we subtracted the jittered population correlation function from the population correlation function based on the residual activity, obtaining the jitter-corrected population correlation function. Corrected peak height and delay was computed by finding the maximum of the jitter-corrected population correlation function, as well as the corresponding delay. Supplementary Fig. 2 illustrates this process.

Comparing interaction structure across time

To determine whether the population activity patterns involved in inter-areal interactions changed during the trial, we leveraged the probabilistic extension of CCA (pCCA)[64]. pCCA is closely related to CCA in that both methods identify the same canonical dimensions. The advantage of pCCA is that it defines an explicit generative model, which we can leverage for model comparison and selection (see Supplementary Information).

Note that the population correlation functions described above could have been computed using pCCA instead of CCA, which would have yielded the same results. We focused there on the first canonical dimension, and did not need the model comparison and selection procedures described below. Thus, solely for clarity of presentation, we opted to introduce the population correlation functions using CCA.

pCCA is defined by the following generative model:

$$\begin{aligned}\mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)\\ \mathbf{x}|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_x\mathbf{z}, \mathbf{\Psi}_x)\\ \mathbf{y}|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_y\mathbf{z}, \mathbf{\Psi}_y)\end{aligned}$$

where $\mathbf{z}$ is a $q \times 1$ latent variable, $\mathbf{x}$ and $\mathbf{y}$ correspond to the neuronal activity recorded in each of two cortical areas, with dimensionalities $p_x \times 1$ and $p_y \times 1$, respectively, p_x and p_y are the number of neurons recorded in each area, and $q \leq \min(p_x, p_y)$. The identity matrix $\mathbf{I}_q$ has dimensions $q \times q$. The mapping matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ have dimensions $p_x \times q$ and $p_y \times q$, respectively. The covariance matrices $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$ have dimensions $p_x \times p_x$ and $p_y \times p_y$, respectively. We assume, without loss of generality, that $\mathbf{x}$ and $\mathbf{y}$ are mean-centered. To fit pCCA, we first applied CCA and used the canonical dimensions and associated canonical correlations to compute the parameters of the pCCA model (see Supplementary Information).

Under the pCCA model, the inter-areal covariance is fully determined by the matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ (see Supplementary Information for an extended discussion of pCCA and its relation to classical CCA). In particular, the column spaces of these matrices define the activity patterns, in each area, along which activity covaries across the two populations.

We used pCCA to compute the $\mathbf{W}_x$ and $\mathbf{W}_y$ matrices at different epochs and compared these matrices, across epochs, to assess whether similar population activity patterns were involved in the inter-areal interaction (Fig. 6). We computed the population activity patterns related across areas (i.e., $\mathbf{W}_x$ and $\mathbf{W}_y$) at one epoch in the trial, and asked how much inter-areal correlation these population activity patterns explained at a different epoch.

Specifically, we first fit a pCCA model with dimensionality q (see procedure below for selecting q) separately for each epoch t, yielding parameters $\theta^t = \{\mathbf{W}_x^t, \mathbf{W}_y^t, \mathbf{\Psi}_x^t, \mathbf{\Psi}_y^t\}$. We then asked: given the observed (sample) within-area covariance matrices at time t, $\mathbf{\Sigma}_{xx}^t$ and $\mathbf{\Sigma}_{yy}^t$, how correlated would the activity across the two areas be if instead of the estimated matrices $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$, the interaction was instead described by the matrices $\mathbf{W}_x^{t'}$ and $\mathbf{W}_y^{t'}$, obtained from a different epoch t'? In other words, how much does the across area correlation change if we compute across-area correlations using population activity patterns defined by $\mathbf{W}_x^{t'}$ and $\mathbf{W}_y^{t'}$, instead of $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$? To quantify the change in correlation, we computed normalized correlations, defined as the total correlation captured at epoch t by the dimensions fit to epoch t' over the total correlation captured by the dimensions fit to epoch t (both computed in a cross-validated manner; see Methods). Misalignment between the column spaces will lead to decreased correlations, and low normalized correlation. On the other hand, if the mapping matrices $\mathbf{W}_x^t$ ($\mathbf{W}_y^t$) and $\mathbf{W}_x^{t'}$ (resp. $\mathbf{W}_y^{t'}$) share the same column space (i.e., if the across-area correlations at epochs t and t' involve the same population activity patterns), the resulting correlations should remain the same, and normalized correlation will close to 1. Algorithm (1) describes this procedure in detail (Supplementary Information).

In order to combine results across recording sessions, in Fig. 6 we used a single value for the latent dimensionality q for all sessions. To select the value of the latent dimensionality q, we first determined the value q^t that maximized the cross-validated data likelihood for each epoch t, in each recording session. For the anesthetized V1-V2 recordings, the average dimensionality across all recording sessions was 3.30 ± 0.09 SEM across epochs in the evoked period and 2.09 ± 0.14 SEM across epochs in the spontaneous period (averages taken across epochs and recording sessions). To avoid comparing spurious canonical dimensions, we choose q to be no greater than both these estimated dimensionalities. Thus, we choose $q = 2$ for these recordings. For the awake V1-V4 recordings the average dimensionality across all recording sessions was 1.8 ± 0.29 SEM across epochs in the evoked period and 2.40 ± 0.24 SEM across epochs in the spontaneous period (averages taken across epochs and recording sessions). Thus, we choose $q = 1$ for these recordings. For both sets of recordings, results were robust to different choices of q, over a reasonable range.

Data availability

V1-V2 data are available at the CRCNS data sharing web site, at `https://doi.org/10.6080/K0B27SHN`. V1-V4 data will be made available upon reasonable request.

Code availability

MATLAB code that supports the data analyses will be made publicly available upon publication.

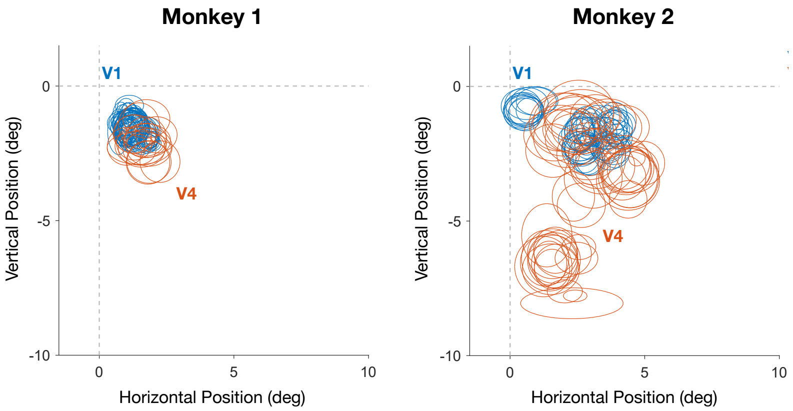
References

61. Jasper, A. I., Tanabe, S. & Kohn, A. Predicting Perceptual Decisions Using Visual Cortical Population Responses and Choice History. *Journal of Neuroscience* **39**, 6714–6727 (2019).

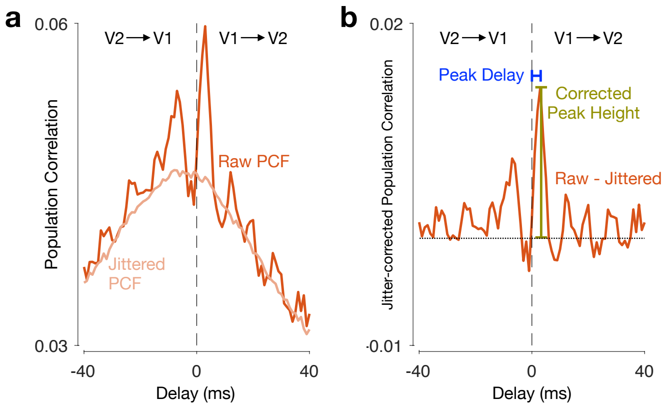
62. Carandini, M. *et al.* Do We Know What the Early Visual System Does? *Journal of Neuroscience* **25**, 10577–10597 (2005).

63. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (1936).

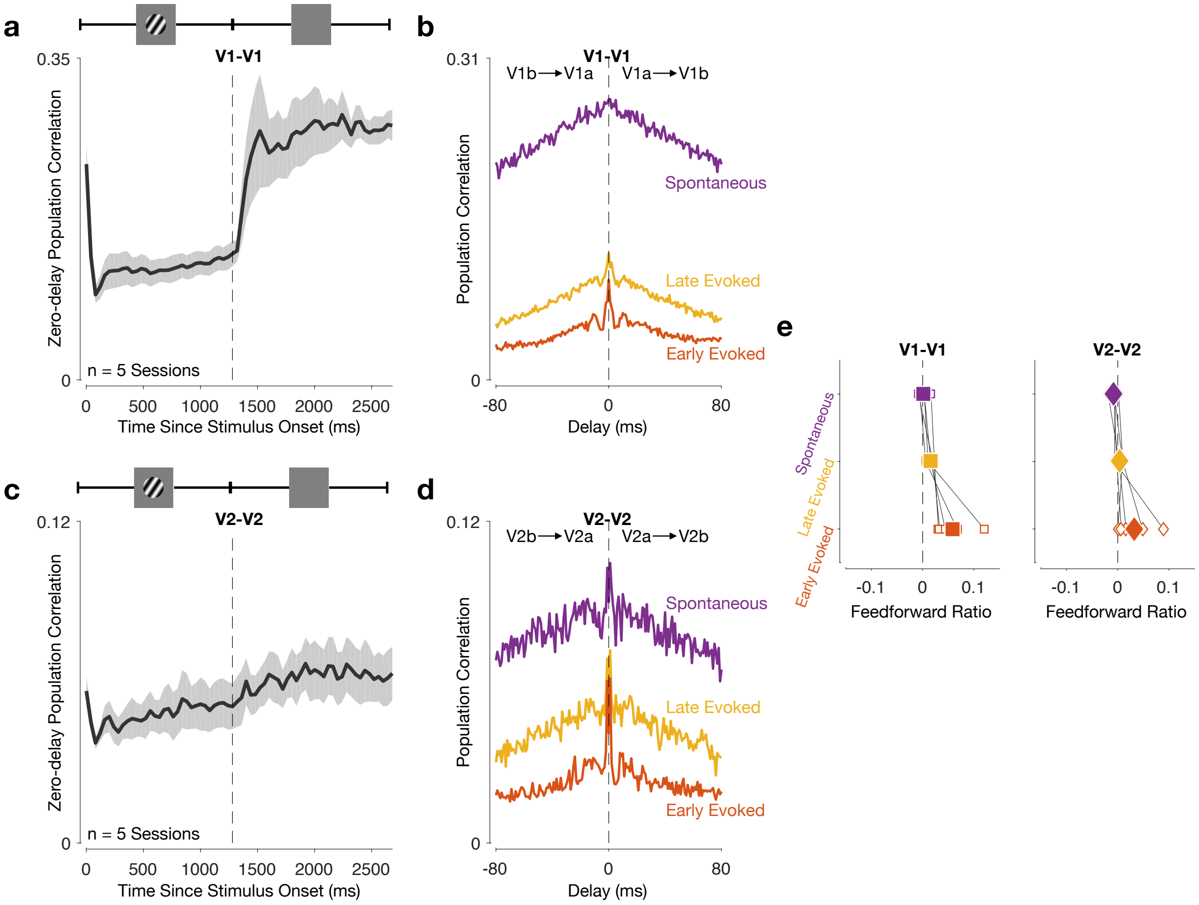
64. Bach, F. R. & Jordan, M. I. A probabilistic interpretation of canonical correlation analysis (2005).



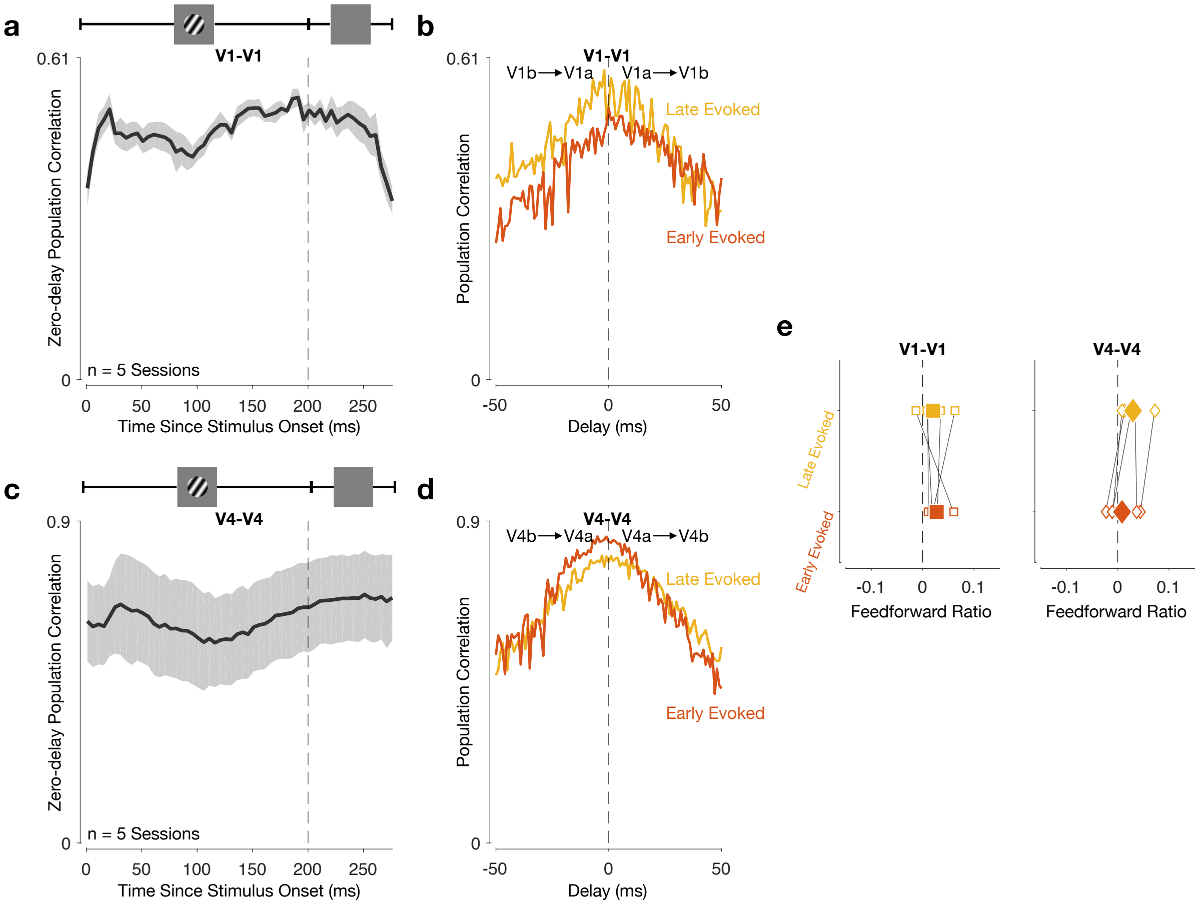
Supplementary Figure 1 Spatial receptive fields for the V1-V4-awake recordings. Lines indicate 60% contour lines of a 2-dimensional Gaussian fit to the receptive fields. Receptive fields were fitted to unsorted multi-unit activity recorded on each channel.



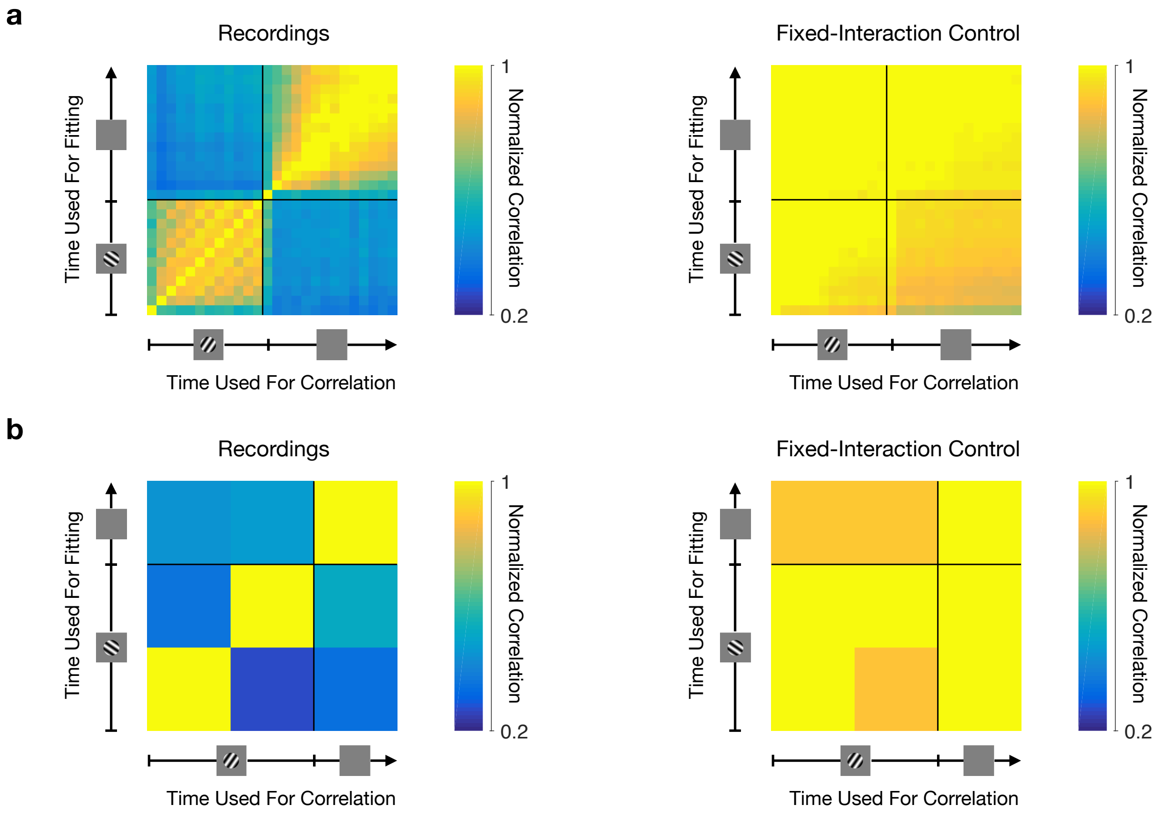
Supplementary Figure 2 Isolanting feedforward peaks using a jitter-corrected population correlation function. **(a)** If a feedforward peak is caused by precise spiking coordination across the two areas, it should still be present after the slow-timescale component of the population correlation function is removed. To remove the slow-timescale component, thereby isolating fast-timescale features in the early evoked activity, we computed a jitter-corrected population function[36]. We first computed a jittered population correlation function (Jittered PCF; 25 ms jitter window), as described in ref. 36. We then obtained a jitter-corrected PCF by subtracting the jittered PCF from the PCF based on residual activity (Raw PCF). **(b)** We computed the peak height by finding the maximum value of the jitter-corrected PCF, as well as the corresponding time delay. In this session, a clear peak can be observed at 3 ms. Results across all recording sessions are shown in Fig. 3e.



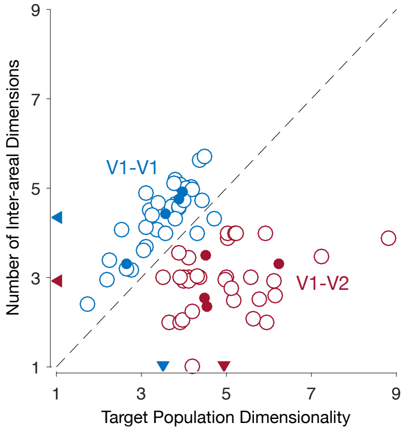
Supplementary Figure 3 Feedforward peak is absent, and feedback-dominated interactions are not evident between subpopulations within V1 or V2. To test whether the effects in Fig. 3 were specific to inter-areal interactions, we randomly divided the neurons in each area into two subpopulations and computed population correlation functions between the subpopulations for each area. **(a)** V1-V1 zero-delay population correlation increases throughout the trial, and is higher for spontaneous activity than for evoked activity. Solid line shows average across all recording sessions. Shading indicates S.E.M. **(b)** V1-V1 population correlation functions for an example session (taken as the average across 10 random divisions into two V1 subpopulations). Same conventions as in Fig. 3b. **(c)** V2-V2 zero-delay population correlation increases throughout the trial, and is higher for spontaneous activity than for evoked activity. Solid line shows average across all recording sessions. Shading indicates S.E.M. **(d)** V2-V2 population correlation function for an example session (taken as the average across 10 random divisions into two V2 subpopulations). Same conventions as in Fig. 3b. **(e)** There are two key features of inter-areal interactions revealed in Fig. 3 which are not present for within-area interactions. First, the feedforward peaks of the population correlation functions for within-area interactions are centered at 0 ms delay (see panels b and d). This is in contrast to the across-area (V1-V2) case, where there is a feedforward peak shortly after stimulus onset (Fig. 3b,e). Second, within-area interactions (V1-V1, left panel; V2-V2, right panel) were neither feedforward- nor feedback-dominated during spontaneous activity (average spontaneous activity feedforward ratio, computed in the -80 to 80 ms delay range: 0.002 ± 0.004 SEM for V1-V1; -0.007 ± 0.003 SEM for V2-V2; t-test for spontaneous activity feedforward ratio, $p = 0.71$ for V1-V1; $p = 0.09$ for V2-V2. This is in contrast to the across-area (V1-V2) case, where interactions were feedback-dominated during spontaneous activity (Fig. 3d). Note that the feedforward ratio is slightly positive for the early evoked period, although the population correlation functions peak at 0 ms time delay throughout the whole trial. This reflects the slightly greater area under the right half compared to the left half of the population correlation function (panels b and d), likely due to the strong change in correlations at stimulus onset (panels a and c). This effect occurs on a slow timescale and motivates our use of jitter-corrected responses reported in the main text (Fig. 3e). Solid symbols show average across all recording sessions, empty symbols correspond to each recording session. Same conventions as in Fig. 3d.



Supplementary Figure 4 Interactions between subpopulations within V1 or V4 recorded in awake animals were neither feedforward- nor feedback-dominated. To test whether the effects in Fig. 4 were specific to inter-areal interactions, we randomly divided the neurons in each area into two subpopulations and computed population correlation functions between the subpopulations for each area. **(a)** V1-V1 zero-delay population correlation is constant throughout the trial. Shading indicates S.E.M. **(b)** V1-V1 population correlation functions for an example session (taken as the average across 25 random divisions into two V1 subpopulations). Same conventions as in Fig. 4b. **(c)** V4-V4 zero-delay population correlation is constant throughout the trial. Solid line shows average across all recording sessions. Shading indicates S.E.M. **(d)** V4-V4 population correlation functions for an example session (taken as the average across 25 random divisions into two V4 subpopulations). Same conventions as in Fig. 4b. **(e)** There are two key features of inter-areal interactions revealed in Fig. 4 which are not present for within-area interactions. First, the feedforward-dominated interaction shortly after stimulus onset (Fig. 4b,e) is absent here, and the correlation functions are centered at 0 ms delay (see panels b and d). Second, the transition from feedforward- to feedback-dominated interactions in the late evoked period (Fig. 4d) is also absent (average late evoked feedforward ratio, computed in the -50 to 50 ms delay range: 0.027 ± 0.008 SEM for V1-V1; 0.031 ± 0.011 SEM for V4-V4; one-sided paired Wilcoxon signed-rank test for difference between early evoked and late evoked activity across all 5 recording sessions, $p = 0.41$ for V1-V1; $p = 0.97$ for V4-V4).



Supplementary Figure 5 Ensuring that changes in activity patterns most related across areas cannot be ascribed to changes in the within-area population covariance structure. **(a)** We generated V1-V2 surrogate data that had approximately the same within-area covariance structure as the recorded data for each epoch, but for which the inter-areal interaction structure was held fixed (see Methods and Supplementary Information). For this synthetic data, our analysis identified a stable interaction structure (right, compare to left reproduced from Fig. 6d which is based on recorded activity). Same conventions as in Fig. 6d. **(b)** The same was true for the V1-V4 interactions (right, compare to left reproduced from Fig. 6e which is based on recorded activity). Same conventions as in Fig. 6e.



Supplementary Figure 6 A communication subspace is evident when using Canonical Correlation Analysis (CCA) to characterize inter-areal interactions. We previously reported that the interaction between V1 and V2 was low dimensional (termed a communication subspace) using Reduced-Rank Regression (RRR)[29]. RRR is closely related to Canonical Correlation Analysis (CCA), which we employed in this work (for a review, see ref. 53, in press). One might wonder whether CCA also identifies a communication subspace between V1 and V2. We repeated the analysis in our previous work, using the same data that was analyzed there[29], but using CCA instead of RRR to relate the activity across areas. To determine the number of dimensions involved in inter-areal interactions, we first evaluated the cross-validated log-likelihood curve for a probabilistic CCA model (pCCA), and picked the number of canonical dimensions that yielded the highest data likelihood. We then fit a pCCA model with the corresponding dimensionality using all trials, and computed the associated inter-areal covariance matrix. Finally, we used Singular Value Decomposition (SVD) to determine the smallest number of dimensions that captured 95% of the inter-areal covariance, and used that number as our estimate of the number inter-areal dimensions.

As in our previous work[29], we found that fewer dimensions were required to characterize inter-areal interactions (V1-V2; red triangle on vertical axis) than within-area interactions

(V1-V1; blue triangle on vertical axis). In contrast to Fig. 3, where we identified a single significant canonical pair for each epoch and time delay, here we identify on average close to 3. This is largely due to the larger binning windows used here (100 ms vs. 1 ms in Fig. 3). Importantly, the lower number of dimensions required to account for inter-areal interactions, compared to within-area interactions, was not a result of lower dimensional activity in the V2 population, as the population activity dimensionality was higher in V2 than in the held-out V1 populations (compare blue and red triangles on horizontal axis). Moreover, the number of predictive dimensions identified by RRR was highly correlated with the number of canonical dimensions identified by CCA (Pearson correlation coefficient $r^2 = 0.89$ across all datasets; not shown). Open circles corresponds to each dataset, filled circles denote mean across datasets for each recording session. Triangles denote mean across all recording sessions.

Supplementary Information

Characterizing changes in the interaction structure

What constitutes a change in the interaction structure? In other words, how can we evaluate whether or not different activity patterns are involved in inter-areal interactions during different trial epochs? Using Canonical Correlation Analysis (CCA, see Methods) to characterize inter-areal interactions, one might wonder whether changes in the canonical dimensions across two epochs are a good indication of a change in the interaction structure. Here, we show that directly levering the canonical dimensions to test for changes in the interaction structure can be misleading, and propose an alternative approach based on the probabilistic CCA (pCCA) model[64].

Suppose we identify q pairs of canonical dimensions, and represent them as the columns of matrices $\mathbf{A}_q$ and $\mathbf{B}_q$, which have dimensions $p_x \times q$ and $p_y \times q$, respectively, where p_x and p_y are the number of recorded neurons in each of the two areas. The column space of each matrix defines a subspace in each area within which activity is most correlated across areas. If one seeks to compare the canonical dimensions identified during two trial epochs, one possibility is to compare the column spaces of matrices $\mathbf{A}_q$ and $\mathbf{B}_q$ for two different epochs.

There is, however, a potential problem with using this approach to ask whether there was a meaningful change in the inter-areal interaction structure. We can illustrate this issue by considering data generated from a pCCA model. pCCA is defined by the following equations:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{1}$$

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}_x\mathbf{z}, \mathbf{\Psi}_x) \tag{2}$$

$$\mathbf{y}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}_y\mathbf{z}, \mathbf{\Psi}_y) \tag{3}$$

where $\mathbf{z}$ is a $q \times 1$ latent variable, $\mathbf{x}$ and $\mathbf{y}$ correspond to the neuronal activity recorded in each of two cortical areas, with dimensionalities $p_x \times 1$ and $p_y \times 1$, respectively, p_x and p_y are the number of neurons recorded in each area, and $q \leq \min(p_x, p_y)$. The identity matrix $\mathbf{I}_q$ has dimensions $q \times q$. The mapping matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ have dimensions $p_x \times q$ and $p_y \times q$, respectively. The covariance matrices $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$ have dimensions $p_x \times p_x$ and $p_y \times p_y$, respectively. We assume, without loss of generality, that $\mathbf{x}$ and $\mathbf{y}$ are mean-centered. CCA and pCCA return the same correlation values, so both methods result in the same population correlation functions. The advantage of pCCA here is that it provides us with a more complete description of the fitted model and its underlying assumptions.

According to pCCA's graphical model (Fig. S1a), we can describe the observed activity in each area as having an "across-area" component and a "within-area" component. The across-area component emerges via the linear mapping between the shared latent variable $\mathbf{z}$ and each observed variable, $\mathbf{x}$ and $\mathbf{y}$. This mapping is defined by the matrices $\mathbf{W}_x$ and $\mathbf{W}_y$. The within-area components are defined to be Gaussian with unconstrained covariance matrices $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$.

The relationship between the column spaces of the matrices $\mathbf{A}_q$ and $\mathbf{B}_q$ computed by classical CCA and the parameters of the pCCA model is given by:

$$\bar{\mathbf{A}}_q = \mathbf{\Sigma}_{xx}^{-1}\mathbf{W}_x \tag{4}$$

$$= \left(\mathbf{W}_x{\mathbf{W}_x}^T + \mathbf{\Psi}_x\right)^{-1}\mathbf{W}_x \tag{5}$$

$$\bar{\mathbf{B}}_q = \mathbf{\Sigma}_{yy}^{-1}\mathbf{W}_y \tag{6}$$

$$= \left(\mathbf{W}_y{\mathbf{W}_y}^T + \mathbf{\Psi}_y\right)^{-1}\mathbf{W}_y \tag{7}$$

where $\bar{\mathbf{A}}_q$ and $\bar{\mathbf{B}}_q$ have the same column space as $\mathbf{A}_q$ and $\mathbf{B}_q$, respectively (see the "Relationship between CCA and pCCA" section below). This shows that the subspaces spanned by the canonical dimensions in each area depend on the within-area noise parameters $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$. Thus, changes to the within-area components lead to changes in the subspaces spanned by the canonical dimensions, even if the across-area components remain fixed. Measuring changes in the interaction structure by measuring to what extent the subspaces spanned by the canonical dimensions differ would thus lead us to conclude that across-area interaction structure had changed, even though only the within-area components were altered.

We can gain further intuition into the pCCA model by inspecting the joint covariance matrix (Fig. S1b) The covariance for each area, $\mathbf{\Sigma}_{xx}$ ($\mathbf{\Sigma}_{yy}$), is composed of an across-area component, $\mathbf{W}_x{\mathbf{W}_x}^T$ (resp. $\mathbf{W}_y{\mathbf{W}_y}^T$) and within-area component, $\mathbf{\Psi}_x$ (resp. $\mathbf{\Psi}_y$). Figure S1c illustrates this covariance decomposition for one of the areas (ellipses represent each covariance component). For the across area covariance, however, we have $\mathbf{\Sigma}_{xy} = \mathbf{W}_x{\mathbf{W}_y}^T = \mathbf{\Sigma}_{yx}^T$. Thus, the across-area covariance structure is solely determined by the linear mapping matrices $\mathbf{W}_x$ and $\mathbf{W}_y$.

Given that the across-area component in the pCCA model is solely determined by the mapping matrices $\mathbf{W}_x$ and $\mathbf{W}_y$, we can quantify changes in the interaction structure by comparing those matrices for different trial epochs. We will take this approach, and use a pCCA model to estimate the $\mathbf{W}_x$ and $\mathbf{W}_y$ matrices, and in turn use changes in these matrices to detect changes in the interaction structure. We need to first define how to measure differences between the $\mathbf{W}_x$ and $\mathbf{W}_y$ matrices estimated at different times during the trial. As mentioned above, $\mathbf{W}_x$ and $\mathbf{W}_y$ are underdetermined, so an element by element comparison (e.g., the Frobenius norm of the difference between two $\mathbf{W}_x$ matrices fit at different epochs in the trial) is not suitable. We defined our difference metric to be based on differences between the column spaces of $\mathbf{W}_x$ and $\mathbf{W}_y$, i.e., our measure of how much the interaction structure changes across different epochs is only sensitive to changes in the subspaces spanned by the dimensions along which activity is related across areas. To be conservative, we will not consider scaling and affine transformations of these dimensions (which do not change the subspace spanned by these dimensions) as changes to the interaction structure, although they might reflect interesting changes for other analysis goals.

Specifically, we will measure differences between the column spaces of $\mathbf{W}_x$ and $\mathbf{W}_y$ by comparing the inter-area correlation these subspaces account for. To compare the interaction structure identified during two epochs in the trial, indexed by t and t', we first fit a pCCA model at each epoch, yielding parameters $\theta^t = \{\mathbf{W}_x^t, \mathbf{W}_y^t, \mathbf{\Psi}_x^t, \mathbf{\Psi}_y^t\}$ and $\theta^{t'} = \{\mathbf{W}_x^{t'}, \mathbf{W}_y^{t'}, \mathbf{\Psi}_x^{t'}, \mathbf{\Psi}_y^{t'}\}$. We then ask: given the observed (sample) within-area covariance matrices at time t', $\mathbf{\Sigma}_{xx}^{t'}$ and $\mathbf{\Sigma}_{yy}^{t'}$, how correlated would the activity across areas be if instead of the estimated matrices $\mathbf{W}_x^{t'}$ and $\mathbf{W}_y^{t'}$, the interaction was instead described by matrices $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$? More specifically, how much do across-area correlations change if we replace the column space of $\mathbf{W}_x^{t'}$ and $\mathbf{W}_y^{t'}$ by the column space of $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$? We can use equations 4 and 6 to compute the subspace spanned by the canonical dimensions induced by $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$ (see the "Relationship between CCA and pCCA" section below):

$$\bar{\mathbf{A}}_q^{t',t} = \boldsymbol{\Sigma}_{xx}^{t'\ -1}\mathbf{W}_x^t \tag{8}$$

$$\bar{\mathbf{B}}_q^{t',t} = \boldsymbol{\Sigma}_{yy}^{t'\ -1}\mathbf{W}_y^t \tag{9}$$

and measure the amount of across-area correlation captured by $\bar{\mathbf{A}}_q^{t',t}$ and $\bar{\mathbf{B}}_q^{t',t}$. We then compare that amount of across-area correlation to the correlation that would have resulted from using $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$ (i.e., the across-area correlation captured by $\bar{\mathbf{A}}_q^{t',t'}$ and $\bar{\mathbf{B}}_q^{t',t'}$). The results of this analysis are shown in Fig. 6. Note that both $\bar{\mathbf{A}}_q^{t',t}$ and $\bar{\mathbf{A}}_q^{t',t'}$ (resp. $\bar{\mathbf{B}}_q^{t',t}$ and $\bar{\mathbf{B}}_q^{t',t'}$) are computed using the same covariance matrix $\boldsymbol{\Sigma}_{xx}^{t'}$ (resp. $\boldsymbol{\Sigma}_{yy}^{t'}$). Thus, any differences between $\bar{\mathbf{A}}_q^{t',t}$ and $\bar{\mathbf{A}}_q^{t',t'}$ (resp. $\bar{\mathbf{B}}_q^{t',t}$ and $\bar{\mathbf{B}}_q^{t',t'}$) are the result of differences between $\mathbf{W}_x^t$ and $\mathbf{W}_x^{t'}$ (resp. $\mathbf{W}_y^t$ and $\mathbf{W}_y^{t'}$). Specifically, differences between the column spaces of $\bar{\mathbf{A}}_q^{t',t}$ and $\bar{\mathbf{A}}_q^{t',t'}$ (resp. $\bar{\mathbf{B}}_q^{t',t}$ and $\bar{\mathbf{B}}_q^{t',t'}$) are the result of differences between the column spaces if $\mathbf{W}_x^t$ and $\mathbf{W}_x^{t'}$ (resp. $\mathbf{W}_y^t$ and $\mathbf{W}_y^{t'}$; see "Relationship between CCA and pCCA" section below). Algorithm (1) describes this process in detail.

Algorithm 1: Comparing inter-area interaction structure across time

```
Result: Normalized correlations $\zeta_q^{t',t}$ for all $t'$ and $t$, and for all choices of $q$

Given sets of observations $\{\mathbf{X}^t, \mathbf{Y}^t\}_{train}$ and $\{\mathbf{X}^t, \mathbf{Y}^t\}_{test}$, for each trial epoch $t$
for $q = 1, ..., \min(p_x, p_y)$ do
    for $\forall t$ do
        Fit pCCA with latent dimensionality $q$ to the training set $\{\mathbf{X}^t, \mathbf{Y}^t\}_{train}$
         yielding: $\theta^t = \{\mathbf{W}_x^t, \mathbf{W}_y^t, \mathbf{\Psi}_x^t, \mathbf{\Psi}_y^t\}$
        for $\forall t'$ do
            Compute across-area correlation in the test set $\{\mathbf{X}^{t'}, \mathbf{Y}^{t'}\}_{test}$:
                1. Compute correlation subspaces $\bar{\mathbf{A}}^{t',t}$ and $\bar{\mathbf{B}}^{t',t}$:
                   $\bar{\mathbf{A}}^{t',t} = {\mathbf{\Sigma}_{xx}^{t'}}^{-1}\mathbf{W}_x^t$ and $\bar{\mathbf{B}}^{t',t} = {\mathbf{\Sigma}_{yy}^{t'}}^{-1}\mathbf{W}_y^t$
                   where $\mathbf{\Sigma}_{xx}^{t'}$ and $\mathbf{\Sigma}_{yy}^{t'}$ are computed using $\{\mathbf{X}^{t'}, \mathbf{Y}^{t'}\}_{test}$
                2. Project the $\{\mathbf{X}^{t'}, \mathbf{Y}^{t'}\}_{test}$ onto $\bar{\mathbf{A}}^{t',t}$ and $\bar{\mathbf{B}}^{t',t}$:
                   $\hat{\mathbf{X}}^{t'} = \mathbf{X}^{t'}\bar{\mathbf{A}}^{t',t}$ and $\hat{\mathbf{Y}}^{t'} = \mathbf{Y}^{t'}\bar{\mathbf{B}}^{t',t}$
                3. Apply CCA to $\{\hat{\mathbf{X}}^{t'}, \hat{\mathbf{Y}}^{t'}\}$ and sum all $q$ canonical correlations,
                   obtaining $r_q^{t',t}$
        end
    end
    for $\forall t, t'$ do
        Compute normalized correlations $\zeta_q^{t',t} = r_q^{t',t} / r_q^{t',t'}$
    end
end
```

Algorithm (1) describes a simple train/test split, but it is easy to generalize this procedure to run within a k-fold cross-validation scheme (and then average the normalized correlations across test folds). In the current study, we employed 10-fold cross-validation.

Fixed interaction structure control

The analysis described in Algorithm (1) was designed to be sensitive only to the column spaces of the $\mathbf{W}_x$ and $\mathbf{W}_y$ matrices. To empirically test that our analysis is insensitive to changes in the remaining pCCA model parameters (i.e., that the changes reported in Fig. 6 are solely due to changes to $\mathbf{W}_x$ and $\mathbf{W}_y$), we devised a control based on the following intuition: if we analyze data where the across-area component is held fixed while the within-area component changes, our method (if it works as we expect it to) should indicate that there is no change in interaction between areas. In other words, if we keep the column spaces constant across epochs, we should find that all normalized correlations will be close to 1 (i.e., we identify the same column spaces throughout the trial). To carry out this control analysis, we generated surrogate data that was as similar as possible to the observed activity (in terms of the first and second order statistics, number of trials and number of observed neurons), but with fixed column spaces for the mapping matrices $\mathbf{W}_x$ and $\mathbf{W}_y$.

To achieve this, we first fit a pCCA model to the recorded neural activity, across all epochs, obtaining matrices $\mathbf{W}_x$ and $\mathbf{W}_y$. We then choose matrices $\hat{\mathbf{\Psi}}_x^t$ and $\hat{\mathbf{\Psi}}_y^t$ for each epoch such that $\mathbf{W}_x{\mathbf{W}_x}^T + \hat{\mathbf{\Psi}}_x^t \approx \mathbf{\Sigma}_{xx}^t$ and $\mathbf{W}_y{\mathbf{W}_y}^T + \hat{\mathbf{\Psi}}_y^t \approx \mathbf{\Sigma}_{yy}^t$ for each epoch t. Note that $\mathbf{W}_x$ and $\mathbf{W}_y$ are the same for all epochs. Figure S2 illustrates this for two epochs t and t'.

For $\hat{\mathbf{\Psi}}_x^t = \mathbf{\Sigma}_{xx}^t - \mathbf{W}_x{\mathbf{W}_x}^T$ to be a proper covariance matrix, it must be positive definite, which is not guaranteed to be the case (similarly for $\hat{\mathbf{\Psi}}_y^t$). A simple way to ensure that Ψ is positive definite is to scale $\mathbf{W}_x$ and $\mathbf{W}_y$ appropriately for each time step, as this operation does not change their column spaces. Algorithm (2) describes the surrogate data generation process in detail.

We found that fixing the column spaces of the $\mathbf{W}_x$ and $\mathbf{W}_y$ in this way led pCCA to identify fixed columns spaces across all epochs (Supplementary Fig. 5a,b), indicating the results in Fig. 6 are not driven by changes in the within-area components but rather by changes in the inter-areal interaction structure.

```
Algorithm 2: Creating surrogate data with a fixed interaction structure
  Result: Surrogate data $\{\hat{\mathbf{X}}^t, \hat{\mathbf{Y}}^t\}$, for each epoch $t$ and for all choices of $q$

  Given the sets of observations, $\{\mathbf{X}^t, \mathbf{Y}^t\}$, for each trial epoch $t$
  for $q = 1, ..., \min(p_x, p_y)$ do
      Fit pCCA with latent dim. $q$ jointly to all sets of observations,
      yielding $\mathbf{W}_x$ and $\mathbf{W}_y$
      for $\forall t$ do
          1. Fit pCCA with latent dim. $q$ to $\{\mathbf{X}^t, \mathbf{Y}^t\}$, yielding $\theta^t = \{\mathbf{W}_x^t, \mathbf{W}_y^t, \mathbf{\Psi}_x^t, \mathbf{\Psi}_y^t\}$

          2. Compute minimum within-area variances ${\sigma_{min}}_x^2$ and ${\sigma_{min}}_y^2$, given by the
             smallest eigenvalues of $\mathbf{\Psi}_x^t$ and $\mathbf{\Psi}_y^t$, respectively

          3. Compute across-area variance ratio, defined as the total across-area
             variance divided by the total variance in each area[1]:
             $\nu_x^t = \text{trace}\left(\mathbf{W}_x^t {\mathbf{W}_x^t}^T\right) / \text{trace}\left(\mathbf{W}_x^t {\mathbf{W}_x^t}^T + \mathbf{\Psi}_x^t\right)$ and
             $\nu_y^t = \text{trace}\left(\mathbf{W}_y^t {\mathbf{W}_y^t}^T\right) / \text{trace}\left(\mathbf{W}_y^t {\mathbf{W}_y^t}^T + \mathbf{\Psi}_y^t\right)$

          4. Scale $\mathbf{W}_x$ and $\mathbf{W}_y$ such that the across-area variance ratios for epoch $t$
             are $\nu_x$ and $\nu_y$, i.e., choose $\alpha_x^t$ and $\alpha_y^t$ such that:
             trace $\left({\alpha_x^t}^2 \mathbf{W}_x {\mathbf{W}_x}^T\right)$ / trace $\left({\alpha_x^t}^2 \mathbf{W}_x {\mathbf{W}_x}^T + \mathbf{\Psi}_x^t\right) = \nu_x^t$ and
             trace $\left({\alpha_y^t}^2 \mathbf{W}_y {\mathbf{W}_y}^T\right)$ / trace $\left({\alpha_y^t}^2 \mathbf{W}_y {\mathbf{W}_y}^T + \mathbf{\Psi}_y^t\right) = \nu_y^t$

          5. Compute $\hat{\mathbf{\Psi}}_x^t = \mathbf{\Sigma}_{xx}^t - {\alpha_x^t}^2 \mathbf{W}_x {\mathbf{W}_x}^T$ and $\hat{\mathbf{\Psi}}_y^t = \mathbf{\Sigma}_{yy}^t - {\alpha_y^t}^2 \mathbf{W}_y {\mathbf{W}_y}^T$

          6. Using the eigenvalue decompositions of $\hat{\mathbf{\Psi}}_x^t$ and $\hat{\mathbf{\Psi}}_y^t$, set their minimum
             variance to ${\sigma_{min}}_x^2$ and ${\sigma_{min}}_y^2$, respectively[2]

          7. Generate surrogate data $\{\hat{\mathbf{X}}^t, \hat{\mathbf{Y}}^t\}$ from a pCCA model with parameters
             $\theta^t = \{{\alpha_x^t}^2 \mathbf{W}_x, {\alpha_y^t}^2 \mathbf{W}_y, \hat{\mathbf{\Psi}}_x^t, \hat{\mathbf{\Psi}}_y^t\}$, with the same number of samples as in
             the entire set of observations, $\{\mathbf{X}^t, \mathbf{Y}^t\}$
      end
  end
```

[2] The scale of the estimated mapping matrices is underdetermined (see "Relationship between CCA and pCCA" section below), so this ratio is underdetermined as well. As an example, scaling $\mathbf{W}_x$ by c and $\mathbf{W}_y$ by $1/c$ (thereby changing ν_x and ν_y) results in an equivalent model, from a data likelihood perspective (provided $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$ remain positive definite). Although we found that keeping the ratio fixed led to good approximations to the covariance matrices $\mathbf{\Sigma}_{xx}^t$ and $\mathbf{\Sigma}_{yy}^t$, this ratio should not be over-interpreted.

[2] This step is required to ensure that $\hat{\mathbf{\Psi}}_x^t$ and $\hat{\mathbf{\Psi}}_y^t$ are positive definite matrices.

Relationship between CCA and pCCA

The correspondence between the classical formulation and the probabilistic variant of CCA was developed by Bach and Jordan[64]. Specifically, they showed that any maximum likelihood solution derived using the probabilistic model (equations 1-3) corresponds to the same set of canonical dimensions identified using classical CCA. In other words, the data likelihood function for pCCA has infinitely many global optima, where all local optima are also global optima, and all global optima correspond to the same set of canonical dimensions. In particular, if we define the top q canonical dimensions identified for each area by classical CCA as $\mathbf{A}_q$ (a $p_x \times q$ matrix) and $\mathbf{B}_q$ (a $p_y \times q$ matrix), the relationship between the canonical dimensions and the linear mapping matrices from pCCA is given by:

$$\mathbf{W}_x = \mathbf{\Sigma}_{xx}\mathbf{A}_q\mathbf{M}_x \tag{10}$$

$$\mathbf{W}_y = \mathbf{\Sigma}_{yy}\mathbf{B}_q\mathbf{M}_y \tag{11}$$

where $\mathbf{M}_x$ and $\mathbf{M}_y$ are arbitrary $q \times q$ matrices such that $\mathbf{M}_x\mathbf{M}_y{}^T = \mathbf{P}_q$ and the spectral norms of $\mathbf{M}_x$ and $\mathbf{M}_y$ are smaller than one. $\mathbf{P}_q$ is a diagonal matrix containing the first q canonical correlations. As an example, $\mathbf{M}_x = \mathbf{M}_y = \mathbf{P}_q^{1/2}$ satisfies these constraints. Any suitable choice of $\mathbf{M}_x$ and $\mathbf{M}_y$ corresponds to a global maximum of the data likelihood. The link between CCA and pCCA is similar to that between PCA and pPCA, and the derivation of this connection largely follows that originally developed for PCA and pPCA[65,66].

In particular, the fact that $\mathbf{M}_x$ and $\mathbf{M}_y$ are underdetermined means that $\mathbf{W}_x$ and $\mathbf{W}_y$ are not uniquely defined when fitting pCCA, i.e., there are many choices of $\mathbf{W}_x$ and $\mathbf{W}_y$ that result in the same canonical dimensions, and maximizing the data likelihood can return any such choices. Importantly, these $\mathbf{W}_x$ ($\mathbf{W}_y$) matrices all have the same column space, i.e., multiplication by $\mathbf{M}_x$ (resp. $\mathbf{M}_y$) does not change the column space of $\mathbf{W}_x$ (resp. $\mathbf{W}_y$).

Given two matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ found by maximizing the data likelihood, we cannot directly compute $\mathbf{A}_q$ and $\mathbf{B}_q$ from these matrices alone, since we don't know which $\mathbf{M}_x$ and $\mathbf{M}_y$ the particular solution we found corresponds to. However, since all the consistent $\mathbf{W}_x$ ($\mathbf{W}_y$) matrices have the same column space, we can find the column spaces of $\mathbf{A}_q$ and $\mathbf{B}_q$ by computing matrices $\bar{\mathbf{A}}_q$ and $\bar{\mathbf{B}}_q$ (equations 4 and 6), since the column space of $\bar{\mathbf{A}}_q$ ($\bar{\mathbf{B}}_q$) is the same as the column space of $\mathbf{A}_q$ (resp. $\mathbf{B}_q$) (see Lemma 1 below). Note that the column space of $\mathbf{A}_q$ (resp. $\mathbf{B}_q$) is the subspace of $\mathbf{x}$ ($\mathbf{y}$) spanned by the canonical dimensions found by classical CCA. The relationship above indicates that the subspace spanned by the canonical dimensions in $\mathbf{A}_q$ ($\mathbf{B}_q$) depends on the column space of $\mathbf{W}_x$ (resp. $\mathbf{W}_y$) and on $\mathbf{\Sigma}_{xx}$ (resp. $\mathbf{\Sigma}_{yy}$). In particular, if $\mathbf{\Sigma}_{xx}$ ($\mathbf{\Sigma}_{yy}$) is held fixed, the column space of $\mathbf{A}_q$ (resp. $\mathbf{B}_q$) is solely determined by the column space of $\mathbf{W}_x$ (resp. $\mathbf{W}_y$; see Lemma 2 below). This observation forms the basis for Algorithm 1, where we ask how well a pCCA model fit to epoch t (yielding $\mathbf{W}_x^t$ and $\mathbf{W}_y^t$) captures correlations at epoch t'.

Lemma 1. *$\bar{\mathbf{A}}_q$ ($\bar{\mathbf{B}}_q$) and $\mathbf{A}_q$ (resp. $\mathbf{B}_q$) have the same column space.*

Proof. We will show that $\bar{\mathbf{A}}_q$ and $\mathbf{A}_q$ have the same column space. The proof for $\bar{\mathbf{B}}_q$ and $\mathbf{B}_q$ is identical. Starting with equation 10:

$$\begin{aligned} \mathbf{W}_x &= \mathbf{\Sigma}_{xx}\mathbf{A}_q\mathbf{M}_x \\ \Leftrightarrow \mathbf{\Sigma}_{xx}^{-1}\mathbf{W}_x &= \mathbf{A}_q\mathbf{M}_x \\ \Leftrightarrow \bar{\mathbf{A}}_q &= \mathbf{A}_q\mathbf{M}_x \end{aligned}$$

where we used the fact that $\mathbf{\Sigma}_{xx}$ is a square positive definite matrix. So long as $\mathbf{M}_x$ is a full rank matrix (i.e., the first q canonical correlations are non-zero), $\bar{\mathbf{A}}_q = \mathbf{A}_q\mathbf{M}_x$ and $\mathbf{A}_q$ have the same column space. □

Lemma 2. *If $\mathbf{\Sigma}_{xx}$ ($\mathbf{\Sigma}_{yy}$) is held fixed, the column space of $\mathbf{A}_q$ (resp. $\mathbf{B}_q$) is solely determined by the column space of $\mathbf{W}_x$ (resp. $\mathbf{W}_y$).*

Proof. We will show that the column space of $\mathbf{A}_q$ depends solely on the column space of $\mathbf{W}_x$ if $\mathbf{\Sigma}_{xx}$ is held fixed. The proof for $\mathbf{B}_q$ is identical. Using the compact singular value decomposition $\mathbf{W}_x = \mathbf{U}\mathbf{D}\mathbf{V}^T$, and inserting it into equation 10:

$$\begin{aligned}
&\mathbf{\Sigma}_{xx}\mathbf{A}_q\mathbf{M}_x = \mathbf{W}_x \\
\Leftrightarrow\, &\mathbf{\Sigma}_{xx}\mathbf{A}_q\mathbf{M}_x = \mathbf{U}\mathbf{D}\mathbf{V}^T \\
\Leftrightarrow\, &\mathbf{A}_q\mathbf{M}_x = \mathbf{\Sigma}_{xx}^{-1}\mathbf{U}\mathbf{D}\mathbf{V}^T \\
\Leftrightarrow\, &\mathbf{A}_q\mathbf{M}_x\mathbf{V}\mathbf{D}^{-1} = \mathbf{\Sigma}_{xx}^{-1}\mathbf{U}
\end{aligned}$$

where we used the fact that $\mathbf{\Sigma}_{xx}$ is a square positive definite matrix. As long as $\mathbf{M}_x$ is a full rank matrix (i.e., the first q canonical correlations are non-zero), $\mathbf{M}_x\mathbf{V}\mathbf{D}^{-1}$ is a square full rank matrix, and thus $\mathbf{A}_q\mathbf{M}_x\mathbf{V}\mathbf{D}^{-1}$ and $\mathbf{A}_q$ have the same column space. So as long as $\mathbf{\Sigma}_{xx}$ is held fixed, the column space of $\mathbf{A}_q$ only depends on $\mathbf{U}$, which is a basis for the column space of $\mathbf{W}_x$. In other words, if we change $\mathbf{W}_x$, only the changes to $\mathbf{U}$ (its column space), and not changes to $\mathbf{D}$ or $\mathbf{V}$, affect the column space of $\mathbf{A}_q$. □

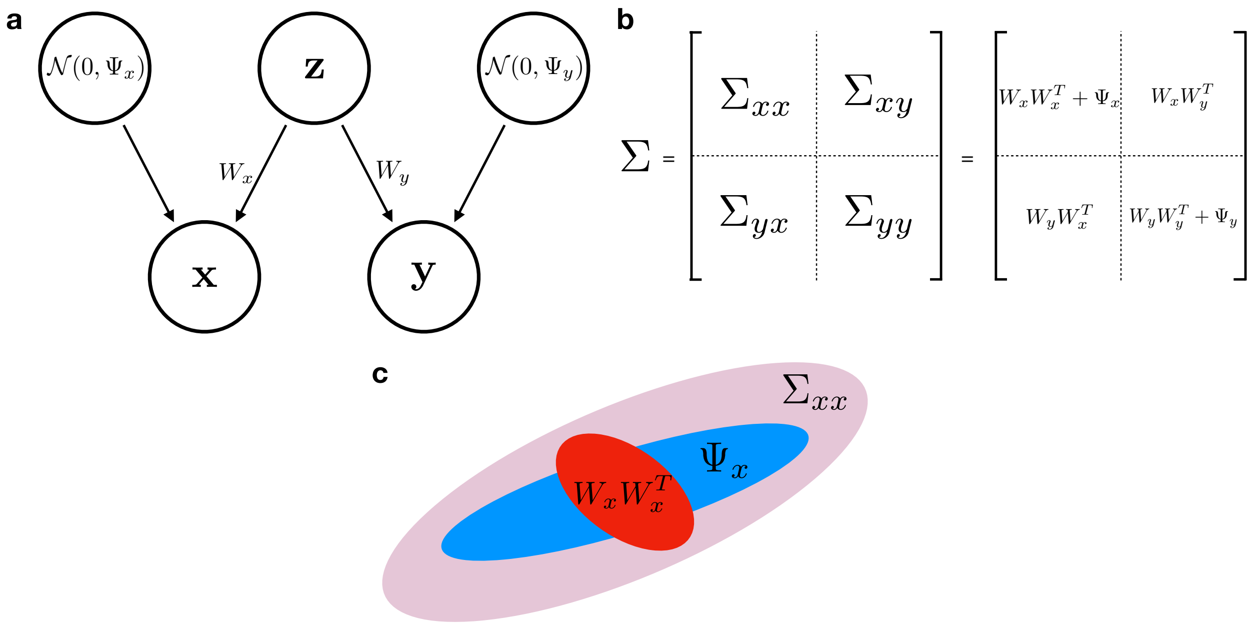

a
$\mathcal{N}(0, \Psi_x)$
$\mathbf{z}$
$\mathcal{N}(0, \Psi_y)$
W_x
W_y
$\mathbf{x}$
$\mathbf{y}$
b
$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} = \begin{bmatrix} W_x W_x^T + \Psi_x & W_x W_y^T \\ W_y W_x^T & W_y W_y^T + \Psi_y \end{bmatrix}$
c
Σ_{xx}
Ψ_x
$W_x W_x^T$

Figure S1 Probabilistic canonical correlation analysis (pCCA) **(a)** pCCA's probabilistic graphical model. **(b)** Summary of the relationship between the data covariance matrices and the pCCA model parameters. **(c)** Graphical representation of the covariance decomposition under a pCCA model for one of the two populations. Red ellipse represents the across-area component; blue ellipse represents the within-area component; pink ellipse represents the total covariance in this area.

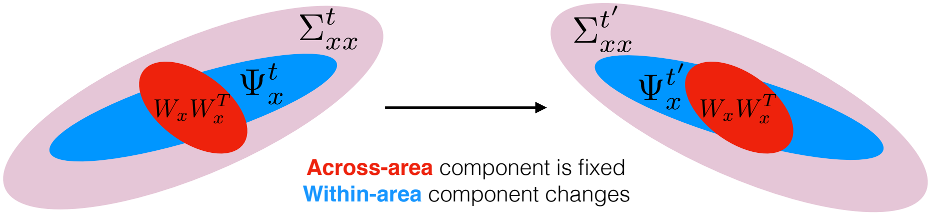


Figure S2 Changing the total covariance in one of the areas while keeping the across-area component fixed. Same conventions as in Fig. S1c.

References

65. Roweis, S. T. EM algorithms for PCA and SPCA. In *Advances in neural information processing systems*, 626–632 (1998).

66. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 611–622 (1999).