HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2014 August 05.

Published in final edited form as:

Nature. 2013 November 7; 503(7474): 78–84. doi:10.1038/nature12742.

Context-dependent computation by recurrent dynamics in prefrontal cortex

Valerio Mante[1,+,*], **David Sussillo**[2,+], **Krishna V. Shenoy**[2,3], and **William T. Newsome**[1]

[1]Howard Hughes Medical Institute and Department of Neurobiology

[2]Department of Electrical Engineering and Neurosciences Program

[3]Departments of Neurobiology and Bioengineering Stanford University, Stanford, CA, United States of America

Summary

Prefrontal cortex is thought to play a fundamental role in flexible, context-dependent behavior, but the exact nature of the computations underlying this role remains largely mysterious. In particular, individual prefrontal neurons often generate remarkably complex responses that defy deep understanding of their contribution to behavior. Here we study prefrontal cortex in monkeys trained to flexibly select and integrate noisy sensory inputs towards a choice. We find that the observed complexity and functional roles of single neurons are readily understood in the framework of a dynamical process unfolding at the level of the population. The population dynamics can be reproduced by a trained recurrent neural network, which suggests a previously unknown mechanism for selection and integration of task-relevant inputs. This mechanism implies that selection and integration are two aspects of a single dynamical process unfolding within the same prefrontal circuits, and potentially provides a novel, general framework for understanding context-dependent computations.

Introduction

Our interactions with the world are inherently flexible. Identical sensory stimuli, for example, can lead to very different behavioral responses depending on ‘context’, which includes goals, prior expectations about upcoming events, and relevant past experiences[1, 2]. Animals can switch rapidly between behavioral contexts, implying the existence of rapid modulation, or ‘gating’, mechanisms within the brain that select relevant sensory information for decision-making and action. A large attention literature suggests that relevant information is selected by top-down modulation of neural activity in early sensory

Correspondence and requests for materials should be addressed to VM (valerio@ini.phys.ethz.ch) and DS (sussillo@stanford.edu).

[+]Equal contribution

[*]Current address: Institute of Neuroinformatics, University of Zurich / ETH Zurich, Zurich, Switzerland

Author contributions

VM and WTN designed the study. VM collected the data. DS implemented the recurrent network. VM and DS analyzed and modeled the data. VM, DS, KVS, and WTN discussed the findings and wrote the paper.

The authors declare no competing financial interests.

areas[3–8], which may take the form of modulation of firing rates[3, 5–7], or modulation of response synchrony within or across areas[4, 5, 8]. The top-down signals underlying such 'early' modulations of sensory activity arise, in part, from prefrontal cortex (PFC)[2, 5], which is known to contribute to representing and maintaining contextual knowledge, ignoring irrelevant information, and suppressing inappropriate actions[1, 2, 9, 10]. These observations have led to the hypothesis that early selection may account for the larger effect of relevant as compared to irrelevant sensory information on contextually-sensitive behavior.

Here we test this hypothesis with a task requiring context-dependent selection and integration of visual stimuli. We trained two macaque monkeys (A and F) to perform two different perceptual discriminations on the same set of visual stimuli (Fig. 1). The monkeys were instructed by a contextual cue to either discriminate the direction of motion or the color of a random-dot display, and to report their choices with a saccade to one of two visual targets (Fig. 1a). While monkeys performed this task, we recorded extracellular responses from neurons in and around the frontal eye field (FEF) (Extended Data Fig. 1a,f), an area of PFC involved in the selection and execution of saccadic eye movements[11, 12], the control of visuo-spatial attention[13], and the integration of information toward visuo-motor decisions[12, 14].

Surprisingly, we find no evidence that irrelevant sensory inputs are gated, or filtered out, prior to the integration stage in PFC, as would be expected from early selection mechanisms[3–8]. Instead, the relevant input appears to be selected late, by the same PFC circuitry that integrates sensory evidence towards a choice. Selection within PFC without prior gating is possible because the representations of the inputs, and of the upcoming choice, are separable at the population level, even though they are deeply entwined at the single neuron level. An appropriately trained recurrent neural network model reproduces key physiological observations and suggests a new mechanism of input selection and integration. The mechanism reflects just two learned features of a dynamical system: an approximate line attractor and a 'selection vector', which are only defined at the level of the population. The model mechanism is readily scalable to large numbers of inputs, suggesting a general solution to the problem of context-dependent computation.

Results

The monkeys successfully discriminated the relevant sensory evidence in each context, while largely ignoring the irrelevant evidence (Fig. 1c-f, monkey A; Extended Data Fig. 2a-d, monkey F). To vary the difficulty of the discrimination, we changed the strength of the motion and color signals randomly from trial to trial (Fig. 1b). In the motion context, the choices of the monkeys depended strongly on the direction of motion of the dots (Fig. 1c), while depending only weakly on color in the same trials (Fig. 1d). The opposite pattern held in the color context: the now relevant color evidence exerted a large effect on choices (Fig. 1f) while motion had only a weak effect (Fig. 1e).

As is common in PFC[1, 2, 15–18], the recorded responses of single neurons appeared to represent several different task-related signals at once, including the monkey's upcoming choice, the context, and the strength of motion and color evidence (Extended Data Figs. 1,3).

Rather than attempting to understand the neural mechanism underlying selective integration by studying the responses of single PFC neurons, we focus on analyzing the responses of the population as a whole. To construct population responses, we pooled data from both single and multi-unit recordings, which yielded equivalent results. The great majority of units were not recorded simultaneously, but rather in separate sessions. Units at all recording locations appeared to contribute to the task-related signals analyzed below (Extended Data Fig. 1) and were thus combined. Overall, we analyzed 388 single-unit and 1014 multi-unit responses from the two monkeys.

To study how the PFC population as a whole dynamically encodes the task variables underlying the monkeys' behavior, we represent population responses as trajectories in neural state space[17, 19–25]. Each point in state space corresponds to a unique pattern of neural activations across the population. Since activations are dynamic, changing over time, the resulting population responses form trajectories in state space.

We focus our analyses on responses in a specific low-dimensional subspace that captures across-trial variance due to the choice of the monkey (choice 1 or 2), the strength and direction of the motion evidence, the strength and direction of the color evidence, and context (motion or color). We estimated this task-related subspace in two steps (Supplementary Information). First, we used principal component analysis (PCA) to obtain an unbiased estimate of the most prominent features (i.e. patterns of activations) in the population response. To 'de-noise' the population responses, we restricted subsequent analyses to the subspace spanned by the first 12 principal components. Second, we used linear regression to define the four orthogonal, task-related axes of choice, motion, color, and context. The projection of the population response onto these axes yields de-mixed estimates of the corresponding task variables, which are mixed both at the level of single neurons (Extended Data Fig. 3) and at the level of individual principal components (Extended Data Fig. 4c,g, see also[26]).

This population analysis yields highly reliable average response trajectories (Fig. 2 and Extended Data Fig. 4q,r) that capture both the temporal dynamics and the relationships among the task variables represented in PFC. In particular, four properties of the population responses provide fundamental constraints on the mechanisms of selection and integration underlying behavior in our task.

First, integration of evidence during dots presentation corresponds to a gradual movement of the population response in state space along the axis of choice (Fig. 2a,f). In both contexts, the trajectories start from a point in state space close to the center of the plots ('dots on', *purple point*), which corresponds to the pattern of population responses at baseline. During the dots presentation the responses then quickly move away from this baseline level, along the axis of choice (*red line*; Fig. 2a,f). Overall, the population response moves in opposite directions on trials corresponding to the two different saccade directions (Fig. 2, *choice 1* vs. *choice 2*). The projection of the population response onto the choice axis (Extended Data Fig. 5b,f) is largely analogous to the 'choice-predictive' signals that have been identified in past studies as approximate integration of evidence during direction discrimination tasks[27].

Second, the sensory inputs into PFC produce patterns of population responses that are very different from those corresponding to either choice, meaning that these signals are separable at the level of the population. Indeed, the population response does not follow straight paths along the choice axis, but instead forms prominent arcs away from it (Figs. 2a,f). The magnitude of each arc along the axes of motion or color reflects the strength of the corresponding sensory evidence (see scale), while its direction (up or down) reflects the sign of the evidence (towards choice 1 or 2, *filled* or *empty* symbols). While the integrated evidence continues to be represented along the axis of choice even after dots offset, the signals along the axes of motion and color are transient—the arcs return to points near the choice axis by the time of dots offset. These signals thus differ from integrated evidence both in terms of the corresponding patterns of activation and in their temporal profile. For these reasons, we interpret them as 'momentary evidence' from the motion and color inputs in favor of the two choices. This interpretation is also consistent with the observed population responses on error trials, for which the momentary evidence points towards the chosen target, but is weaker than on correct trials (Extended Data Fig. 5c,d; *red curves*).

Third, context appears to have no substantial effect on the direction of the axes of choice, motion, and color, and only weak effects on the strength of the signals represented along these axes. When estimated separately during the motion and color contexts, the two resulting sets of axes span largely overlapping subspaces (see Supplementary Information, Table 1); thus a single set of three axes (the *red*, *black*, and *blue axes* in Fig. 2a-f, estimated by combining trials across contexts) is sufficient to capture the effects of choice, motion and color on the population responses in either context. A comparison of the population responses across contexts (Fig. 2a-c vs. d-f) reveals that a single, stable activity pattern is responsible for integrating the relevant evidence in both contexts (the choice axis), while similarly stable activity patterns represent the momentary motion and color evidence in both contexts (motion and color axes). Surprisingly, motion and color inputs result in comparable deflections along the motion and color axes, respectively, whether they are relevant or not (compare Fig. 2a to 2d, and 2f to 2c).

Fourth, while the directions of the axes of choice, motion, and color are largely invariant with context, their location in state space is not. The responses during the motion and color contexts occupy different parts of state space, and the corresponding trajectories are well separated along the axis of context (Extended Data Fig. 6a,b).

These properties of the population responses, which are summarized schematically in Fig. 3a, can be compared to the predictions of current models of context-dependent selection and integration (Fig. 3b-d). Here we first focus on three fundamentally different mechanisms of selection that could each explain why the motion input, for example, influences choices in the motion context (Fig. 3, *top row*) but not in the color context (Fig. 3, *bottom row*). In the framework of our task the three models predict population responses that differ substantially from each other (Fig. 3b-d), and can thus be validated or rejected by our PFC recordings (Fig. 3a).

The first model (Fig. 3b) is based on two widely accepted hypotheses about the mechanisms underlying selection and integration of evidence. First, it assumes that inputs are selected

early[3–8], such that a given input drives PFC responses when relevant (*gray arrow* in Fig. 3b *top*), but is filtered out before reaching PFC when irrelevant (no *gray arrow* in Fig. 3b, *bottom*). Second, it assumes that the relevant input directly elicits a pattern of activation in PFC resembling those corresponding to a choice (the *gray arrow* in Fig. 3b, *top*, points along the axis of choice), as would be expected by current models of integration[28, 29].

Both hypotheses are difficult to reconcile with the recorded PFC responses. While the strength of each input is somewhat reduced when it is irrelevant compared to when it is relevant, the magnitude of the observed reduction seems too small to account for the behavioral effects. For instance, irrelevant motion of high coherence (Fig. 2d *black*), elicits a larger deflection along the motion axis (relative to baseline, *magenta dot*) than relevant motion of intermediate coherence (Fig. 2a, *dark gray*). Yet the former has almost no behavioral effect (Fig. 1e) while the latter has a large behavioral effect (Fig. 1c). The analogous observation holds for the color input (Fig. 2c,f and Fig. 1d,f), strongly suggesting that the magnitude of the momentary evidence alone does not determine whether the corresponding input is integrated. Furthermore, the actual momentary motion input is represented along a direction that has little overlap with the choice axis, resulting in curved trajectories (Fig. 3a) that differ markedly from the straight trajectories predicted by the early selection model (Fig. 3b).

The observed PFC responses also rule out two additional models of selection presented in Fig. 3. In the absence of early selection, a motion input might be selected within PFC by modifying the angle between the choice and motion axes (i.e. the similarity between patterns of neural activity representing choice and momentary motion evidence) across contexts. This angle could be modified either by changing the direction of the motion axis between contexts while keeping the choice axis fixed (Fig. 3c), or vice versa (Fig. 3d). In both cases, the motion input would elicit movement of the population along the axis of choice in the motion context (top row), but not in the color context (bottom row) since the motion and choice axes have little or no overlap in the color context. At the single neuron level, variable axes that change direction across contexts would be reflected as complex, nonlinear interactions between context and the other task-variables, which have been proposed in some task-switching models[30, 31]. However, our data (Fig. 2, Fig. 3a) lend little support for variable choice (Fig. 3d) or input (Fig. 3c) axes. More generally, the PFC data from monkey A rule out any model of integration for which the degree of overlap between the direction of the momentary evidence and the axis of choice determines how much the corresponding input affects behavior.

The representation of task variables in PFC of monkey F replicates all but one key feature observed in monkey A. Most importantly, population responses along the choice and motion axes (Extended Data Fig. 7a,d) closely match those observed in monkey A (Fig. 2a,d); thus physiological data from both monkeys are consistent in rejecting current models of selection and integration of motion inputs (Fig. 3b-d). The color signal in monkey F, however, is equivocal. On the one hand, the representation of the color input closely resembles that of a choice (Extended Data Fig. 1g,i), as expected from the early selection model described above (Fig. 3b). On the other hand, the color input is also weakly represented along the color axis in both contexts (vertical displacement of trajectories, Extended Data Fig. 7c,f). For the

color input in monkey F, therefore, we cannot with confidence accept or reject the early selection model. Finally, as in monkey A, context is represented in monkey F along a separate axis of context (Extended Data Fig. 6c,d).

In summary, the population responses in both monkeys are difficult to reconcile with current models of selection and integration (see also Extended Data Fig. 8). Rather, the selective integration of the motion input in monkeys A and F, and of the color input in monkey A, must rely on a mechanism for which the very same input into PFC leads to movement along a fixed axis of choice in one context but not another.

To identify such a mechanism, we trained a network of recurrently connected, nonlinear neurons[32] to solve a task analogous to the one solved by the monkeys (Fig. 4). Notably, we only defined 'what' the network should do, with minimal constraints on 'how' it should do it[32–34]. Thus the solution achieved by the network is not hand-built into the network architecture. On each trial, neurons in the network receive two independent sensory inputs that mimic the momentary evidence for motion and color in a single random dot stimulus. The network also receives a contextual input that mimics the contextual signal provided to the monkeys, instructing the network to discriminate either the motion or the color input. The network activity is read out by a single linear readout, corresponding to a weighted sum over the responses of all neurons in the network (see Supplementary Information). As in PFC, the contextual input does not affect the strength of the sensory inputs—selection occurs within the same network that integrates evidence toward a decision.

We trained the network[35] to make a binary choice on each trial—an output of +1 at the end of the stimulus presentation if the relevant evidence pointed leftward, or a −1 if it pointed rightward. After training, the model qualitatively reproduces the monkeys' behavior, confirming that the model solves the selection problem at the 'behavioral' level (Extended Data Fig. 2e-h).

We first analyzed model population trajectories in the subspace spanned by the axes of choice, motion, and color, and found that they reproduce the four main features of the PFC population responses discussed above (Fig. 5 and Extended Data Fig. 9a-g). First, integration of evidence corresponds to gradual movement of the population response along the choice axis. Second, momentary motion and color evidence 'push' the population away from the choice axis, resulting in trajectories that are parametrically ordered along the motion and color axes. Third, the direction of the axes of choice, motion, and color are largely invariant with context, as are the strength of the motion and color inputs, since these are not gated before entering the network. Fourth, the trajectories during motion and color contexts are separated along the axis of context (Extended Data Fig. 9f,g). Model and physiological dynamics differ strikingly in one respect—signals along the input axes are transient in the physiology, but not in the model, yielding PFC trajectories that curve back to the choice axis before the end of the viewing interval (compare Fig. 5a,f to Fig. 2a,f). This difference suggests that the sensory inputs to PFC are attenuated after a decision is reached. Additional differences between the model and the physiological dynamics can be readily explained by previously proposed imperfections in the evidence integration process, such as 'urgency' signals[36, 37] or instability in the integrator[38] (Extended Data Fig. 10).

We then 'reversed engineered' the model[33] to discover its mechanism of selective integration. The global features of the model activity are easily explained by the overall arrangement of fixed points of the dynamics[33] (Fig. 5), which result from the synaptic connectivity learned during training. Fixed points (*red crosses*) correspond to patterns of neuronal activations (i.e., locations in state space) that are stable when the sensory inputs are turned off. First, we find that the model generates a multitude of fixed points, which are approximately arranged to form two lines along the choice axis. The two sets of fixed points are separated along the axis of context (Extended Data Fig. 9f,g) and never exist together—one exists in the motion context (Fig. 5a-c), the other in the color context (Fig. 5d-f). Second, the responses around each fixed point are approximately stable only along a single dimension pointing towards the neighboring fixed points (*red lines*), while responses along any other dimension rapidly collapse back to the fixed points. Therefore, each set of fixed points approximates a line attractor[39]. Finally, two stable attractors (*large red crosses*), corresponding to the two possible choices, delimit each line attractor.

The integration of the relevant evidence is thus implemented in the model as movement along an approximate line attractor[39]. The model population response, however, does not move strictly along the line attractor. Like the physiological data, model trajectories move parallel to the line attractors (the choice axis) at a distance proportional to the average strength of the sensory inputs, reflecting the momentary sensory evidence (Fig. 5a,c,d,f). After the inputs are turned off (Fig. 5, *purple data points*), the responses rapidly relax back to the line attractor.

To understand how the relevant input is selected for integration along a line attractor, we analyzed the local dynamics of model responses around the identified fixed points[33] (Fig. 6). To simplify the analysis, we studied how the model responds to brief pulses of motion or color inputs (Fig. 6a), rather than the noisy, temporally extended inputs used above. Before a pulse, we initialize the state of the network to one of the identified fixed points (Fig. 6a, *red crosses*). Locally around a fixed point, the responses of the full, nonlinear model can then be approximated by a linear dynamical system (see Supplementary Information), whose dynamics can be more easily understood[33].

Both the motion and color inputs (i.e. the corresponding pulses) have substantial projections onto the line attractor (Fig. 6a) but, crucially, the size of these projections does not predict the extent to which each input will be integrated. For instance, in both contexts the motion pulses have similar projections onto the line attractor (Fig. 6a, *left panels*), and yet they result in large movement along the attractor in the motion context (*top*) but not in the color context (*bottom*).

The selection of the inputs instead relies on context-dependent relaxation of the network dynamics after the end of the pulse, which reverses movement along the line attractor caused by the irrelevant pulse (Fig. 6a, *top right* and *bottom left*) and enhances the effects of the relevant pulse (Fig. 6a, *top left* and *bottom right*). These relaxation dynamics, while counterintuitive, nevertheless follow a very simple rule. For a given context, the relaxation always occurs on a path that is orthogonal to a specific direction in state space, which we call the 'selection vector' (Fig. 6b). The direction of the selection vector, like the direction

of the line attractor, is a property of the recurrent synaptic weights learned by the model during training (see Supplementary Information). Unlike the line attractor, however, the orientation of the selection vector changes with context—it projects strongly onto the relevant input, but is orthogonal to the irrelevant one (Fig. 6b). As a consequence, the relaxation dynamics around the line attractor are context-dependent. This mechanism explains how the same sensory input can result in movement along the line attractor in one context but not the other (Fig. 6b).

The line attractor and the selection vector are sufficient to explain the linearized dynamics around each fixed point (see Supplementary Information), and approximate well the responses of the full model (*magenta curves*, Fig. 6a). Concretely, the line attractor and the selection vector correspond to the right and left zero-eigenvector of the underlying linear system. Within a context, these locally defined eigenvectors point in a remarkably consistent direction across different fixed points—the selection vector, in particular, is always parallel to the relevant input and orthogonal to the irrelevant input (Fig. 6c and Extended Data Fig. 10q-s). As a result, the two line attractors (Fig. 6c) exhibit relaxation dynamics appropriate for selecting the relevant input along their entire length.

Discussion

We describe a novel mechanism underlying flexible, context-dependent selection of sensory inputs and their integration towards a choice (see[39–41] for related concepts). This mechanism is sufficient to explain the selection and integration of motion inputs in both monkeys, and of color inputs in monkey A, which are not filtered out by context before they reach PFC.

A randomly initialized, recurrent neural network trained to solve a task analogous to the monkeys' task reproduces the main features of the data, and analysis of the trained network elucidates the novel selection mechanism. Integration along line attractors, and its relation to the selection vector, has been described before[39]. However, our model demonstrates for the first time how a single non-linear model can implement flexible computations by reconfiguring the selection vector and the corresponding recurrent dynamics based on a contextual input. Counter-intuitively, in the model the projection of an input onto the line attractor does not determine the extent to which it is integrated, a manifestation of 'non-normal' dynamics[40, 42, 43] (see Supplementary Information).

Our results show that the modulation of sensory responses is not necessary to select among sensory inputs (see also[44–46]). Consistent with this conclusion, two studies employing tasks similar to ours[47, 48], as well as our own recordings in area MT of monkey A (data not shown), have found no evidence for consistent firing rate modulations in the relevant sensory areas. The dynamical process outlined in this paper is fully sufficient for context-dependent selection in a variety of behavioral paradigms[3–8], but it need not be exclusive. Multiple selection mechanisms may exist within the brain.

In summary, our results suggest that computations in prefrontal cortex emerge from the concerted dynamics of large populations of neurons, and are best studied in the framework of dynamical systems[17, 19–23, 39, 49]. Remarkably, the rich dynamics of PFC responses

during selection and integration of inputs can be characterized and understood with just two features of a dynamical system—the line attractor and the selection vector, which are defined only at the level of the neural population. This parsimonious account of cortical dynamics contrasts strikingly with the complexity of single neuron responses typically observed in PFC and other integrative structures, which reveal multiplexed representation of many task-relevant and choice-related signals[1, 2, 15, 16, 25, 50]. In light of our results, these mixtures of signals can be interpreted as separable representations at the level of the neural population[15, 17, 25]. A fundamental function of PFC may be to generate such separable representations, and to flexibly link them through appropriate recurrent dynamics to generate the desired behavioral outputs.

Methods

Methods are provided in the Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jessica Powell, Sania Fong, and Julian Brown for technical assistance, and Larry Abbott, Lubert Stryer, Sonja Hohl, Surya Ganguli, Maneesh Sahani, Roozbeh Kiani, Chris Moore and Tanmoy Bhattacharya for helpful discussions. VM and WTN were supported by HHMI and the Air Force Research Laboratory (FA9550-07-1-0537), DS and KVS by an NIH Director's Pioneer Award (1DP1OD006409) and DARPA REPAIR (N66001-10-C-2010).

References

1. Fuster, JM. The prefrontal cortex. 4th edn. Academic Press; 2008.
2. Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 2001; 24:167–202. [PubMed: 11283309]
3. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Annu Rev Neurosci. 1995; 18:193–222. [PubMed: 7605061]
4. Schroeder CE, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 2009; 32:9–18. [PubMed: 19012975]
5. Noudoost B, Chang MH, Steinmetz NA, Moore T. Top-down control of visual attention. Curr Opin Neurobiol. 2010; 20:183–190. [PubMed: 20303256]
6. Reynolds JH, Chelazzi L. Attentional modulation of visual processing. Annu Rev Neurosci. 2004; 27:611–647. [PubMed: 15217345]
7. Maunsell JH, Treue S. Feature-based attention in visual cortex. Trends Neurosci. 2006; 29:317–322. [PubMed: 16697058]
8. Fries P. Neuronal gamma-band synchronization as a fundamental process in cortical computation. Annu Rev Neurosci. 2009; 32:209–224. [PubMed: 19400723]
9. Mansouri FA, Tanaka K, Buckley MJ. Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. Nat Rev Neurosci. 2009; 10:141–152. [PubMed: 19153577]
10. Tanji J, Hoshi E. Role of the lateral prefrontal cortex in executive behavioral control. Physiol Rev. 2008; 88:37–57. [PubMed: 18195082]
11. Bruce CJ, Goldberg ME. Primate frontal eye fields. I. Single neurons discharging before saccades. J Neurophysiol. 1985; 53:603–635. [PubMed: 3981231]

12. Schall JD. The neural selection and control of saccades by the frontal eye field. Philos Trans R Soc Lond B Biol Sci. 2002; 357:1073–1082. [PubMed: 12217175]
13. Moore T. The neurobiology of visual attention: finding sources. Curr Opin Neurobiol. 2006; 16:159–165. [PubMed: 16563729]
14. Kim JN, Shadlen MN. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. Nat Neurosci. 1999; 2:176–185. [PubMed: 10195203]
15. Machens CK, Romo R, Brody CD. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. J Neurosci. 2010; 30:350–360. [PubMed: 20053916]
16. Rigotti M, et al. The importance of mixed selectivity in complex cognitive tasks. Nature. 2013; 497:585–590. [PubMed: 23685452]
17. Stokes MG, et al. Dynamic coding for cognitive control in prefrontal cortex. Neuron. 2013; 78:364–375. [PubMed: 23562541]
18. Hernandez A, et al. Decoding a perceptual decision process across cortex. Neuron. 2010; 66:300–314. [PubMed: 20435005]
19. Churchland MM, et al. Neural population dynamics during reaching. Nature. 2012; 487:51–56. [PubMed: 22722855]
20. Shenoy KV, Sahani M, Churchland MM. Cortical control of arm movements: a dynamical systems perspective. Annu Rev Neurosci. 2013; 36:337–359. [PubMed: 23725001]
21. Stopfer M, Jayaraman V, Laurent G. Intensity versus identity coding in an olfactory system. Neuron. 2003; 39:991–1004. [PubMed: 12971898]
22. Briggman KL, Abarbanel HD, Kristan WB Jr. Optical imaging of neuronal populations during decision-making. Science. 2005; 307:896–901. [PubMed: 15705844]
23. Harvey CD, Coen P, Tank DW. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature. 2012; 484:62–68. [PubMed: 22419153]
24. Afshar A, et al. Single-trial neural correlates of arm movement preparation. Neuron. 2011; 71:555–564. [PubMed: 21835350]
25. Sigala N, Kusunoki M, Nimmo-Smith I, Gaffan D, Duncan J. Hierarchical coding for sequential task events in the monkey prefrontal cortex. Proc Natl Acad Sci U S A. 2008; 105:11969–11974. [PubMed: 18689686]
26. Machens CK. Demixing population activity in higher cortical areas. Front Comput Neurosci. 2010; 4:126. [PubMed: 21031029]
27. Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol. 2001; 86:1916–1936. [PubMed: 11600651]
28. Mazurek ME, Roitman JD, Ditterich J, Shadlen MN. A role for neural integrators in perceptual decision making. Cereb Cortex. 2003; 13:1257–1269. [PubMed: 14576217]
29. Wang XJ. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 2002; 36:955–968. [PubMed: 12467598]
30. Cohen JD, Dunbar K, McClelland JL. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev. 1990; 97:332–361. [PubMed: 2200075]
31. Deco G, Rolls ET. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. Eur J Neurosci. 2003; 18:2374–2390. [PubMed: 14622200]
32. Sussillo D, Abbott LF. Generating coherent patterns of activity from chaotic neural networks. Neuron. 2009; 63:544–557. [PubMed: 19709635]
33. Sussillo D, Barak O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Comput. 2013; 25:626–649. [PubMed: 23272922]
34. Zipser D, Andersen RA. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature. 1988; 331:679–684. [PubMed: 3344044]
35. Martens, J.; Sutskever, I. Learning recurrent neural networks with hessian-free optimization; Proceedings of the 28th International Conference on Machine Learning; 2011.
36. Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. Nat Neurosci. 2008; 11:693–702. [PubMed: 18488024]

37. Reddi BA, Carpenter RH. The influence of urgency on decision time. Nat Neurosci. 2000; 3:827–830. [PubMed: 10903577]
38. Brunton BW, Botvinick MM, Brody CD. Rats and humans can optimally accumulate evidence for decision-making. Science. 2013; 340:95–98. [PubMed: 23559254]
39. Seung HS. How the brain keeps the eyes still. Proc Natl Acad Sci U S A. 1996; 93:13339–13344. [PubMed: 8917592]
40. Goldman MS. Memory without feedback in a neural network. Neuron. 2009; 61:621–634. [PubMed: 19249281]
41. Sejnowski TJ. On the stochastic dynamics of neuronal interaction. Biol Cybern. 1976; 22:203–211. [PubMed: 953078]
42. Murphy BK, Miller KD. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. Neuron. 2009; 61:635–648. [PubMed: 19249282]
43. Ganguli S, Huh D, Sompolinsky H. Memory traces in dynamical systems. Proc Natl Acad Sci U S A. 2008; 105:18970–18975. [PubMed: 19020074]
44. Salinas E. Context-dependent selection of visuomotor maps. BMC Neurosci. 2004; 5:47. [PubMed: 15563737]
45. Zenon A, Krauzlis RJ. Attention deficits without cortical neuronal deficits. Nature. 2012; 489:434–437. [PubMed: 22972195]
46. Roy JE, Riesenhuber M, Poggio T, Miller EK. Prefrontal cortex activity during flexible categorization. J Neurosci. 2010; 30:8519–8528. [PubMed: 20573899]
47. Sasaki R, Uka T. Dynamic readout of behaviorally relevant signals from area MT during task switching. Neuron. 2009; 62:147–157. [PubMed: 19376074]
48. Katzner S, Busse L, Treue S. Attention to the Color of a Moving Stimulus Modulates Motion-Signal Processing in Macaque Area MT: Evidence for a Unified Attentional System. Front Syst Neurosci. 2009; 3:12. [PubMed: 19893762]
49. Machens CK, Romo R, Brody CD. Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science. 2005; 307:1121–1124. [PubMed: 15718474]
50. Huk AC, Meister ML. Neural correlates and neural computations in posterior parietal cortex during perceptual decision-making. Front Integr Neurosci. 2012; 6:86. [PubMed: 23087623]

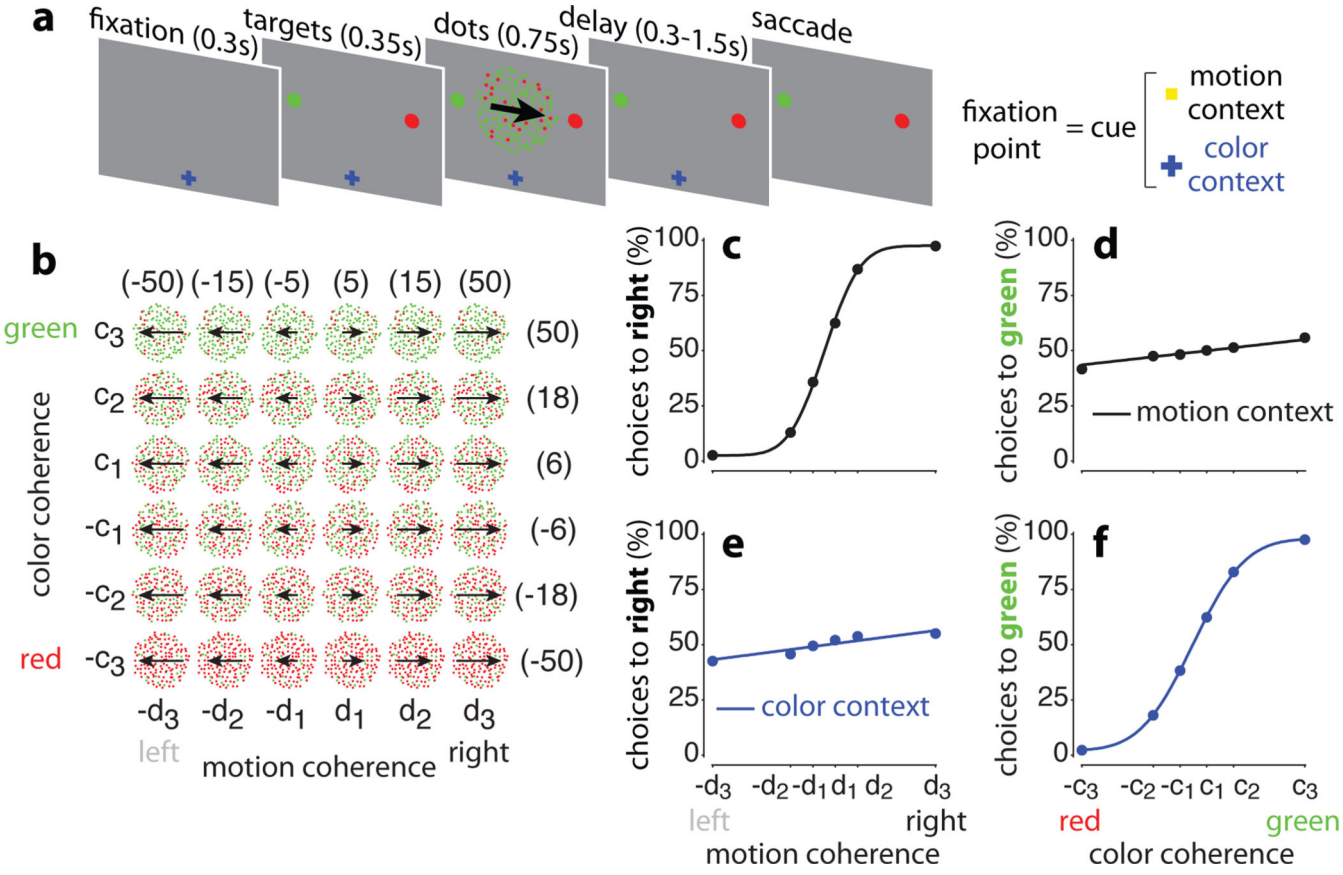


Figure 1. Behavioral task and psychophysical performance

a, Task structure. Monkeys were instructed by a contextual cue to either discriminate the motion or the color of a random-dot stimulus, and indicate their choice with a saccade to one of two targets. Depending on context, monkeys were rewarded for choosing the target matching the prevalent direction of motion (*motion context*) or the prevalent color (*color context*) of the random-dots. Context was indicated by the shape and color of the fixation point; offset of the fixation point was the 'go cue', signaling the monkey to indicate its choice via the operant saccade. **b**, Stimulus set. The motion and color coherence of the dots was chosen randomly on each trial. We slightly varied the coherence values on each day, to equate performance across contexts and sessions (numbers in parentheses: average coherences (%) across sessions for monkey A). **c-f**, Psychophysical performance for monkey A in the motion (*top*) and color contexts (*bottom*), averaged over 80 recording sessions (163,187 trials). Performance is shown as a function of motion (*left*) or color (*right*) coherence in each behavioral context. The curves are fits of a behavioral model.

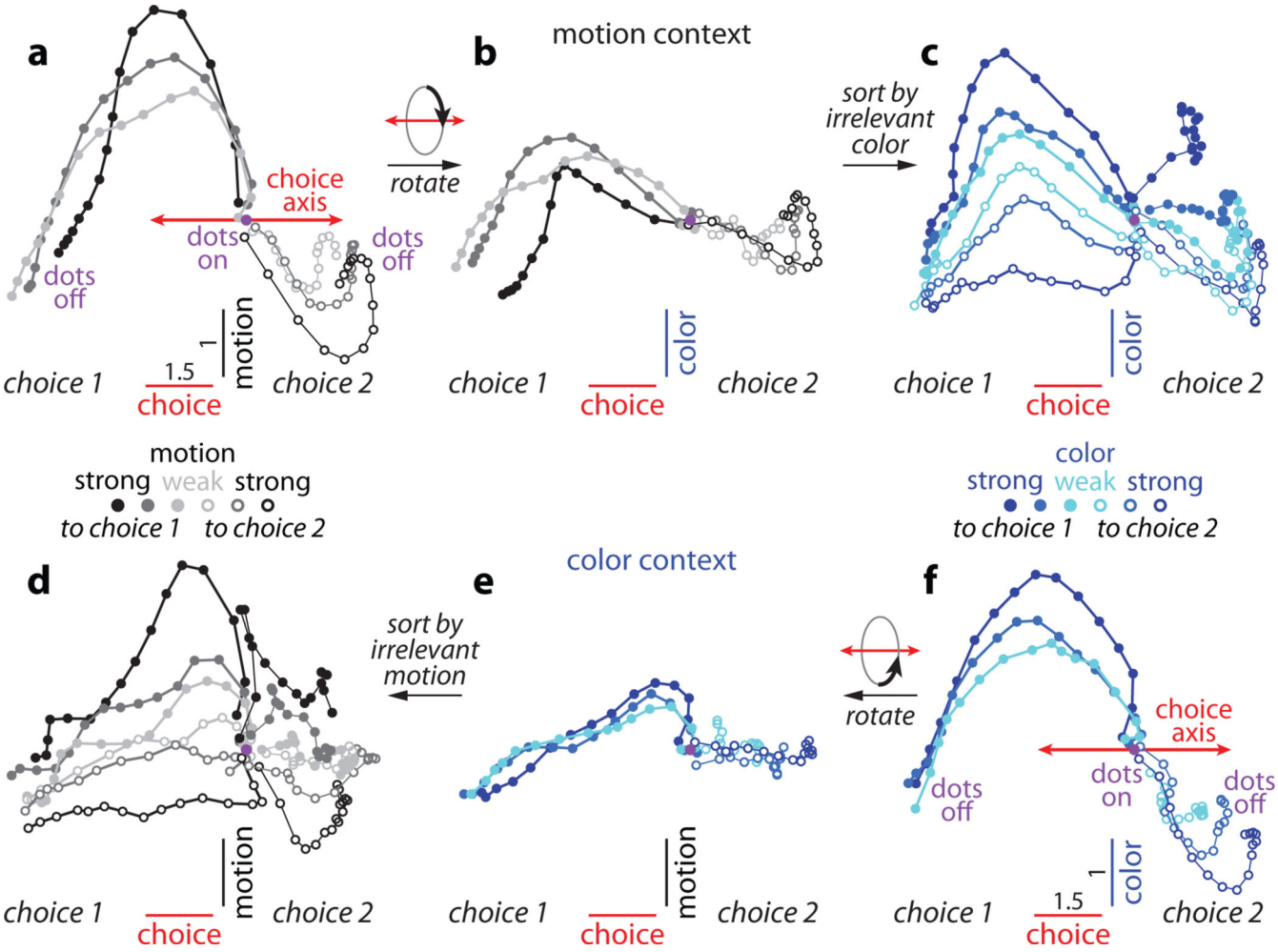


Figure 2. Dynamics of population responses in PFC

The average population response for a given condition and time is represented as a point in state space. Responses from correct trials only are shown from 100ms after dots onset (*dots on*, *purple circle*) to 100ms after dots offset (*dots off*) in 50ms steps, and are projected into the three-dimensional subspace capturing the variance due to the monkey's choice (along the *choice* axis), and to the direction and strength of the motion (*motion* axis) and color (*color* axis) inputs. Units are arbitrary; components along the motion and color axes are enhanced relative to the choice axis (see scale bars in **a,f**). Conditions (see *color bars*) are defined based on context (*motion context*, top; *color context*, bottom), on the location of the chosen target (*choice 1* vs. *choice 2*) and either on the direction and strength of the motion (*gray colors*) or the color input (*blue colors*). Here, choice 1 corresponds to the target in the response field of the recorded neurons. The direction of the color input does not refer to the color of the dots *per se* (red or green), but to whether the color points towards choice 1 or choice 2 (see Supplementary Information, section 6.4, for a detailed description of the conditions). **a**, Effect of choice and the relevant motion input in the motion context, projected onto the axes of choice and motion. **b**, Same data as in **a**, but rotated by 90° around the axis of choice to reveal the projection onto the axis of color. **c**, Same trials as in **b**, but re-sorted according to the direction and strength of the irrelevant color input. **d-f**,

responses in the color context, analogous to **a-c**. Responses are averaged to show the effects of the relevant color (**e,f**) or the irrelevant motion input (**d**). For relevant inputs (**a,b** and **e,f**), correct choices occur only when the sensory stimulus points towards the chosen target (3 conditions per chosen target); for irrelevant inputs (**c,d**), however, the stimulus can point either towards or away from the chosen target on correct trials (6 conditions per chosen target).

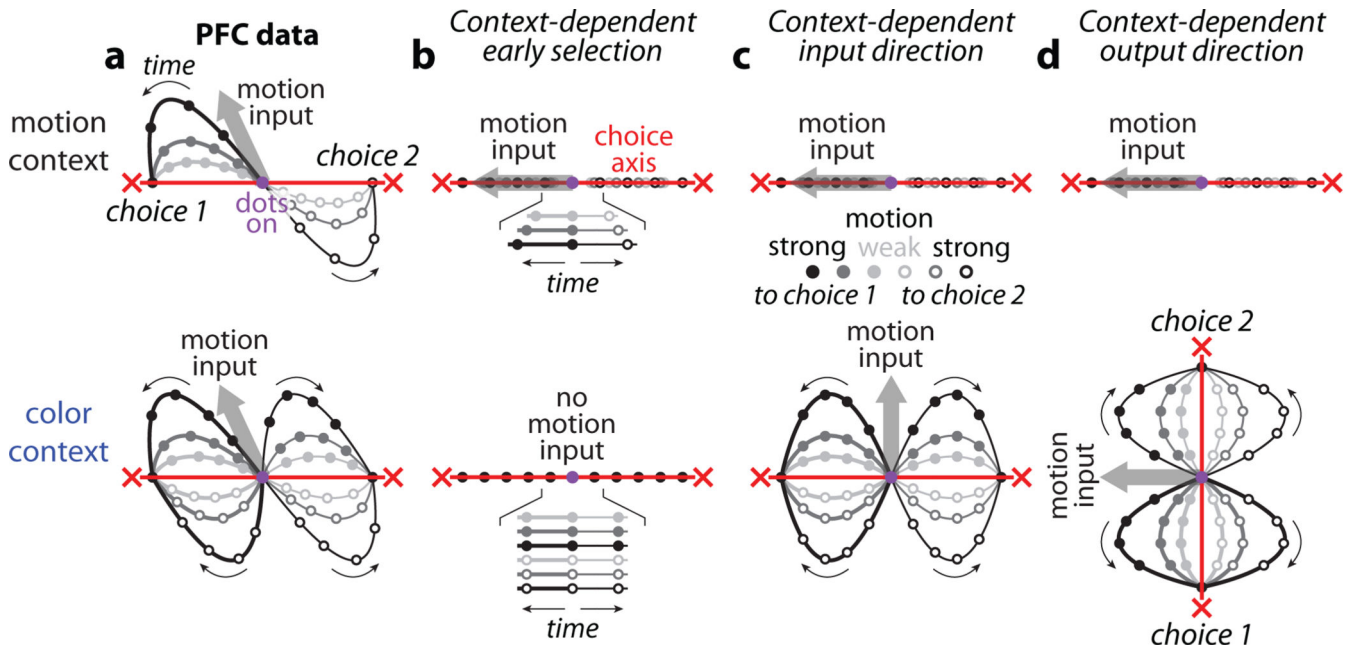


Figure 3. Models of selective integration inconsistent with PFC responses

Schematic representation of population responses observed in PFC (**a**) and expected by several models of selective integration (**b-d**). The models differ from the PFC responses with respect to the relative directions and context dependence of the choice axis (*red lines*) and the inputs (*thick gray arrows*; only motion input shown). The relevant input is integrated as movement along the choice axis towards one of two choices (*red crosses*). A motion input towards choice 1 'pushes' the responses along the direction of the *gray arrow* (towards choice 2: opposite direction). Same conditions as in Fig. 2a (motion context, *top*) and Fig. 2d (color context, *bottom*). As in Fig. 2a and d, a single two-dimensional subspace (which contains the choice axis and motion input) is used to represent responses from both contexts. **a**, Idealized schematic of the actual PFC trajectories shown in Fig. 2a,d. Both the choice axis and motion input are stable between contexts. The motion input pushes the population response away from the choice axis. **b**, Early selection model. When relevant (*top*), the motion input pushes the population response along the choice axis. When irrelevant (*bottom*), the motion input is filtered out before reaching PFC (*no thick gray arrow*) and thus exerts no effect on choice. All trajectories fall on top of each other in both contexts, but the rate of movement along the choice axis increases with motion strength only in the motion context (*insets* show enlarged trajectories distributed vertically for clarity). **c**, Context dependent input direction. Motion input direction varies between contexts, while the choice axis is stable. Inputs are not filtered out before PFC; rather they are selected based on their projection onto the choice axis. **d**, Context-dependent output direction. Similar selection mechanism to **c**, except that the choice axis varies between contexts, while the motion input is stable. The effects of the motion input on PFC responses in both monkeys (schematized in **a**) and the effects of the color input in monkey A are inconsistent with predictions of the three models in **b-d** (respectively, Fig. 2a,d; Extended Data Fig. 7a,d; Fig. 2f,c).

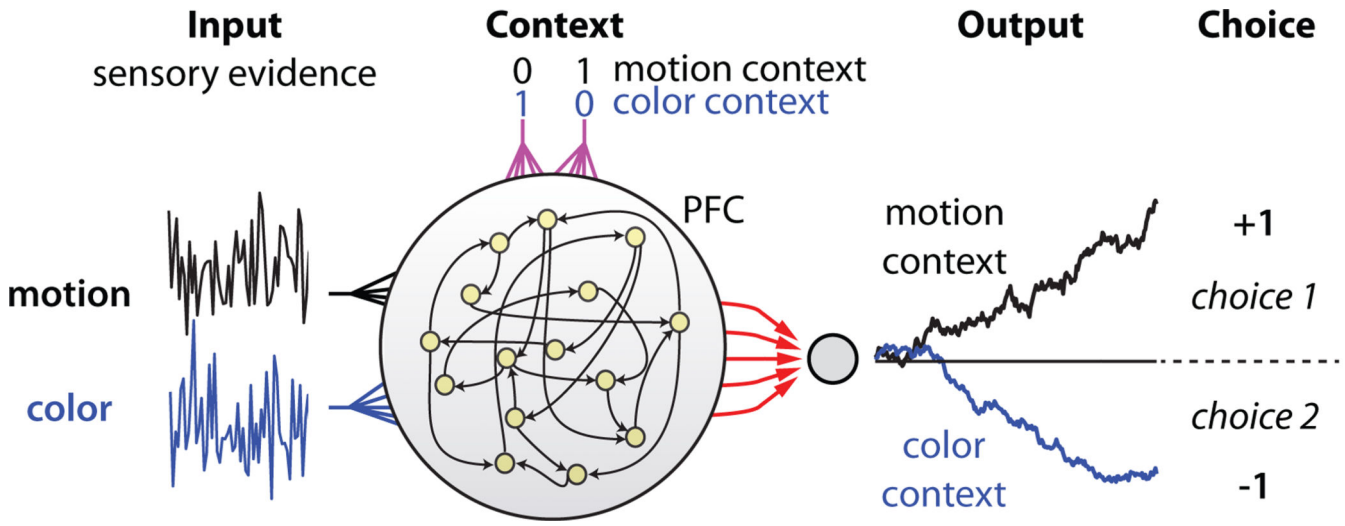


Figure 4. A neural network model of input selection and integration

PFC is modeled as a network of recurrently connected, nonlinear, rate neurons that receive independent motion, color, and contextual inputs. The network is fully recurrently connected, and each unit receives both motion and color inputs as well as two inputs that indicate context. At each time step, the sensory inputs are drawn from two normal distributions whose means correspond to the average strengths of the motion and color evidence on a given trial. The contextual inputs take one of two values (0 or 1), which instruct the network to discriminate either the motion or the color input. The network is read out by a single linear readout, corresponding to a weighted sum over the responses of all neurons (*red arrows*). We trained the network (with back-propagation[35]) to make a binary choice, i.e. to generate an output of +1 at the end of the stimulus presentation if the relevant evidence pointed towards choice 1, or a −1 if it pointed towards choice 2. Before training, all synaptic strengths were randomly initialized.

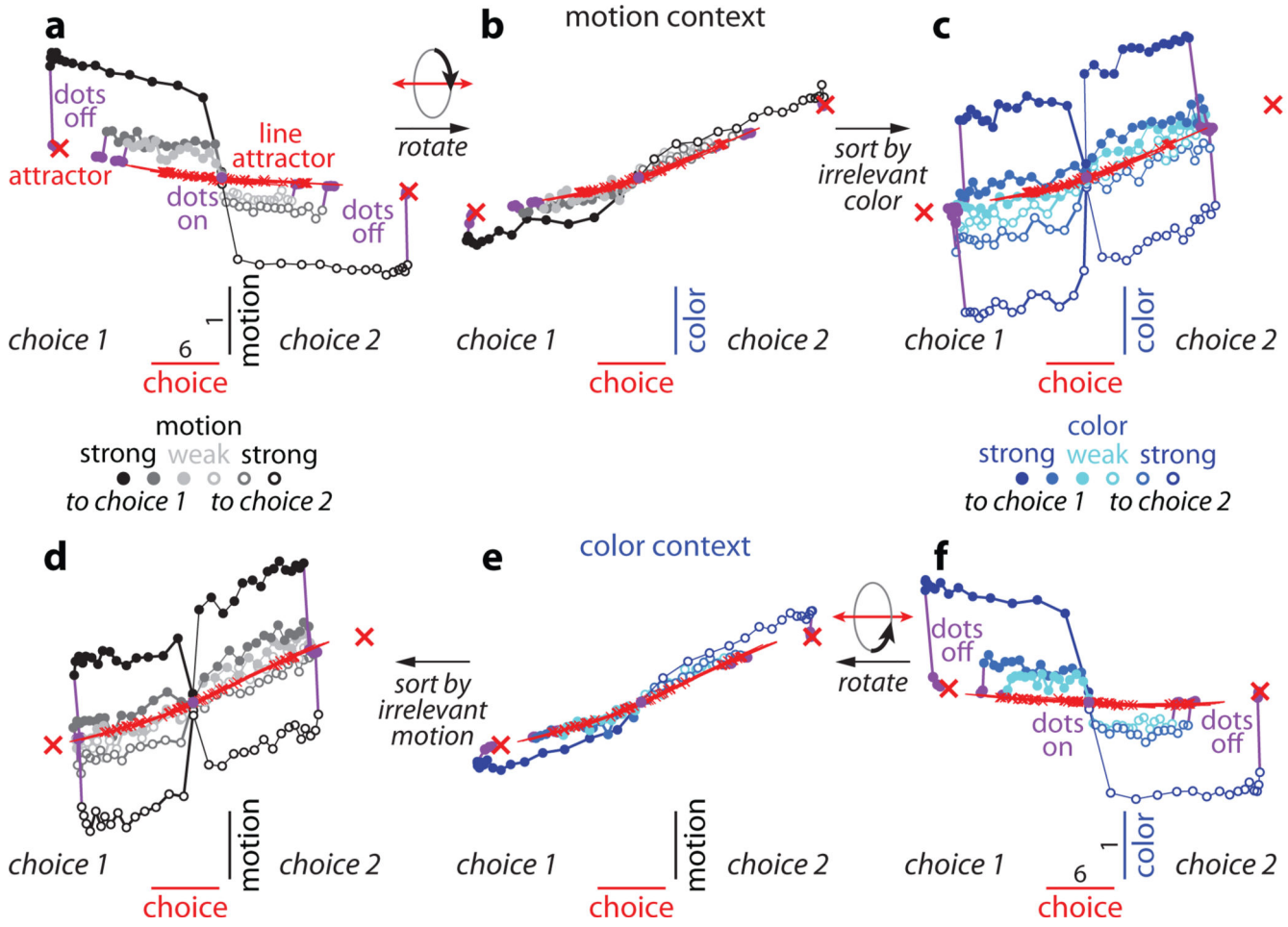


Figure 5. Model dynamics and fixed points analysis

a-f, Dynamics of model population responses, same conventions as in Fig. 2. Responses are projected into the three-dimensional subspace spanned by the axes of choice, motion, and color (defined here based on the model synaptic weights, see Supplementary Information, section 7.6). Movement along the choice axis corresponds to integration of evidence, and the motion and color inputs deflect the trajectories along the corresponding input axes. Fixed points of the dynamics (*red crosses*) were computed separately for motion (**a-c**) and color contexts (**d-f**) in the absence of sensory inputs (see Supplementary Information, section 7.5). The fixed points are 'marginally stable' (i.e. one eigenvalue of the linearized dynamics is close to zero, while all others have strongly negative real parts; see Supplementary Information). The locally computed right zero-eigenvectors (*red lines*) point to the neighboring fixed points, which thus approximate a line attractor in each context. After the inputs are turned off (*dots off*, *purple* data points and lines) the responses relax back towards the line attractor. Each line attractor ends in two 'stable' attractors (i.e. all eigenvalues have strongly negative real parts, *large crosses*) corresponding to model outputs of +1 and −1 (i.e. choice 1 or 2).

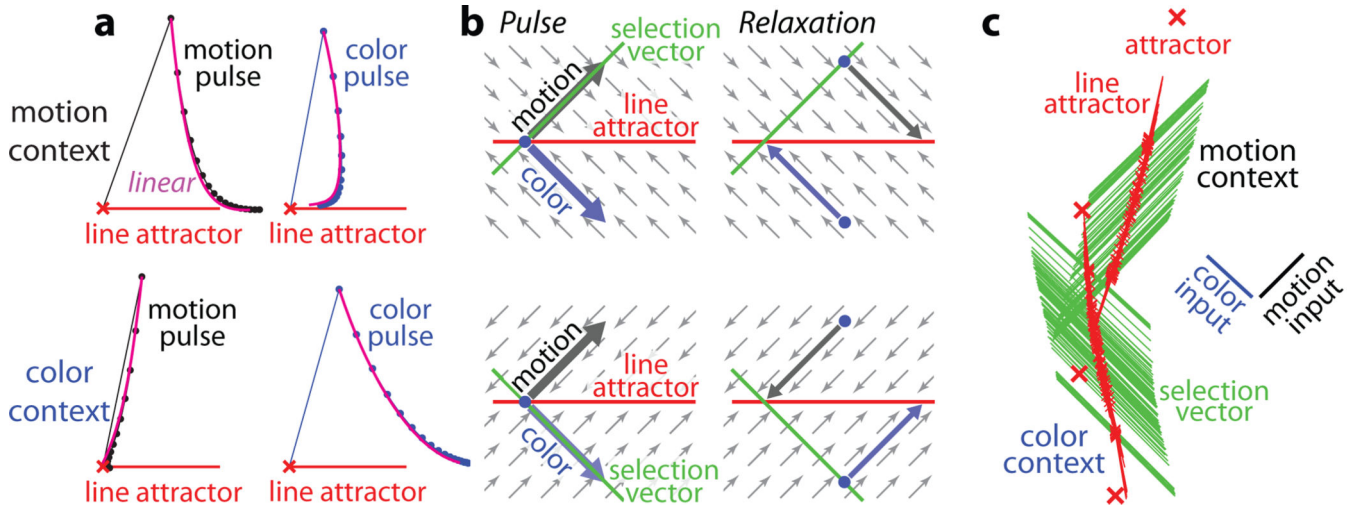
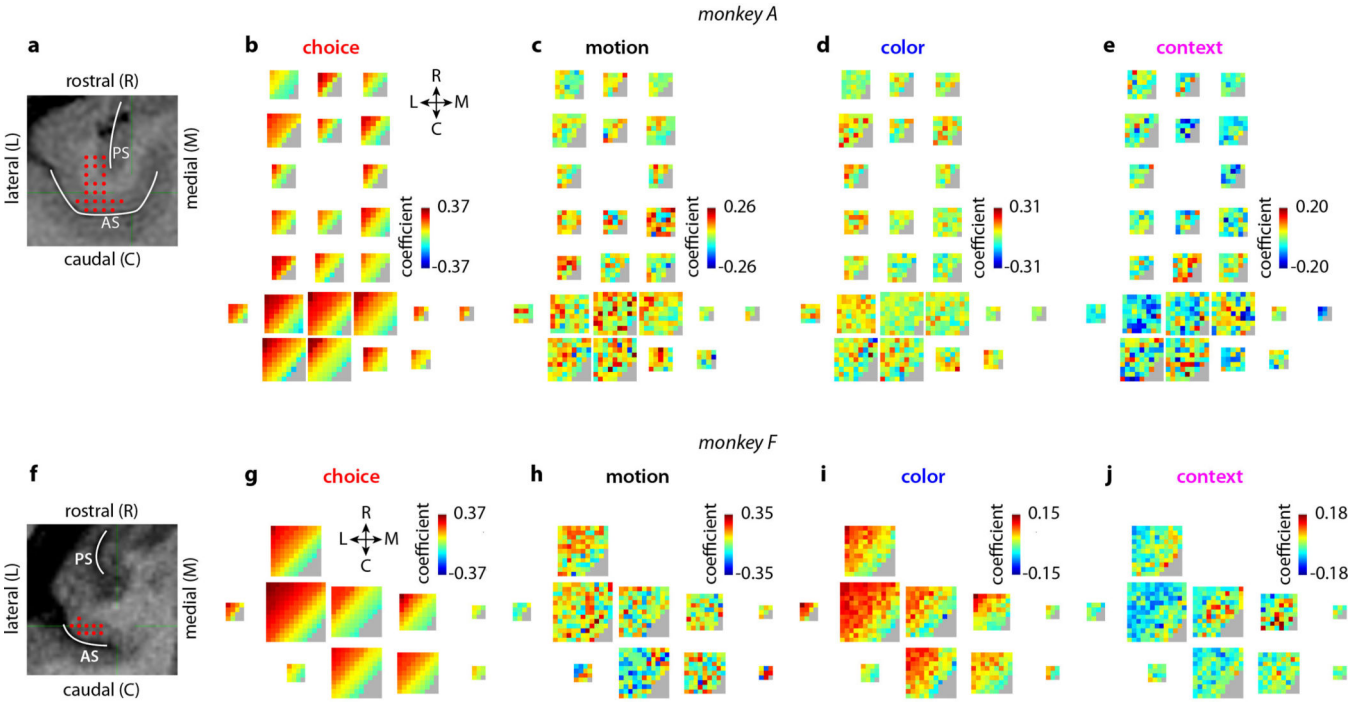


Figure 6. Selection and integration by recurrent dynamics

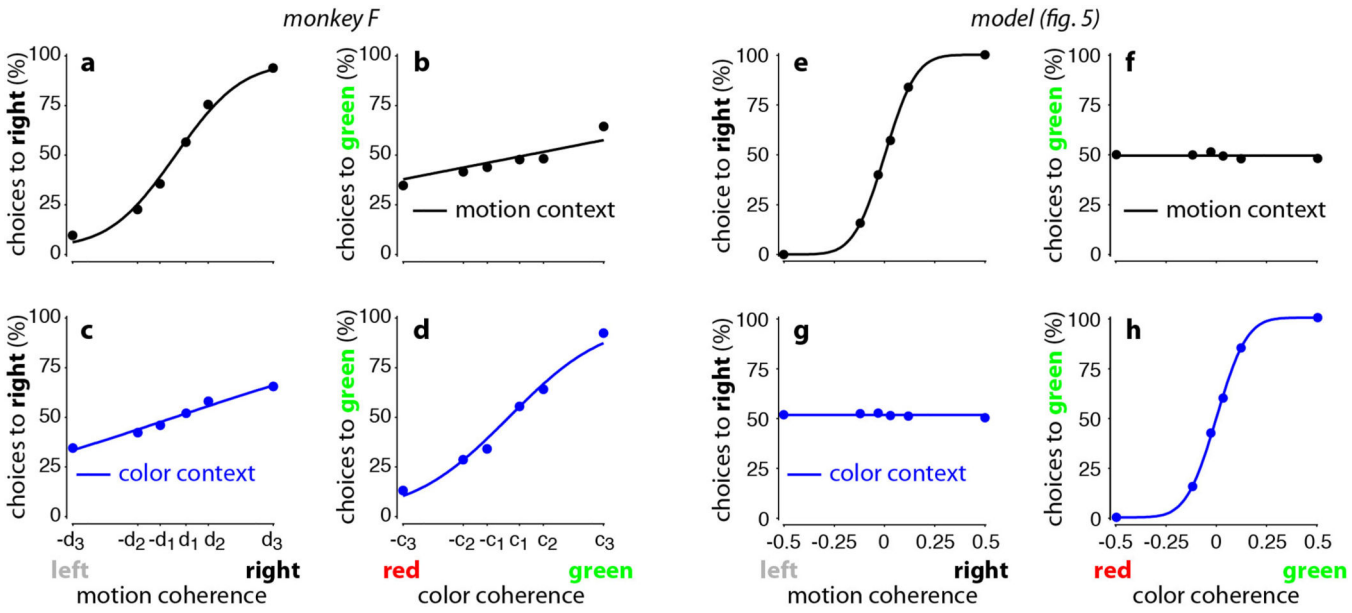
a, Average model population response to short (1ms) pulses of motion (*left*) and color inputs (*right*) during motion (*top*) and color contexts (*bottom*). Motion or color inputs (*solid lines*) are initiated when the system is steady at one of the identified fixed points (*red crosses*), and subsequent relaxation back to the line attractor is simulated (dots: 3ms intervals) and averaged across fixed points. The size of the pulses approximately corresponds to the length of the scale bars in Fig. 5. Selection of the relevant input results from the context-dependent relaxation of the recurrent dynamics after the pulse, and is well approximated by the linearized dynamics around the fixed points (*magenta lines*). Responses are projected into the two-dimensional subspace spanned by the direction of the pulse and the locally computed line attractor (the right zero-eigenvector of the linearized dynamics). **b**, Explanation of how the same input pulse (*left*) leads to evidence integration in one context, but is ignored in the other (*right*). Relaxation towards the line attractor (*small arrows*) is always orthogonal to the context-dependent selection vector, and reverses the effects of the irrelevant pulse. **c**, Global arrangement of the line attractor (*red*) and selection vector (*green*) at each fixed point. Inputs are selected by the selection vector, which is orthogonal to the contextually irrelevant input (note input axes, *right*), and integrated along the line attractor.



Extended Data Figure 1. Recording locations and task-related patterns of population activity in PFC

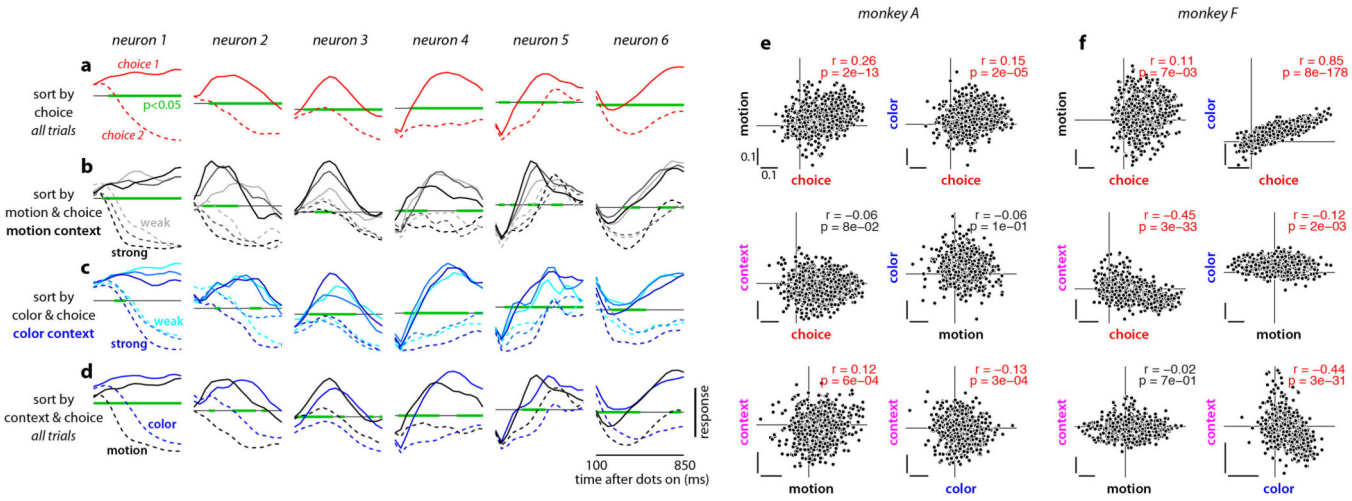
a, Recording locations (*red dots*) in monkey A are shown on anatomical magnetic resonance images in imaging planes that were oriented perpendicularly to the direction of electrode penetrations. Electrodes were lowered through a grid (1mm spacing) positioned over the arcuate sulcus (AS). Recordings covered the entire depth of the AS and extended rostrally onto the prearcuate gyrus and cortex near and lateral to the principal sulcus. **b-e**, Representation of 4 task variables in the population response. Each multi-colored square corresponds to a recording location (*red dots*) in **a**. Within each square, each pixel corresponds to a unit recorded from that grid position, such that each square represents all the units recorded at the corresponding location. The color of a pixel indicates the de-noised regression coefficient of choice (**b**), motion coherence (**c**), color coherence (**d**), and context (**e**) for a given unit (color bars; gray: no units). These coefficients describe how much the trial-by-trial firing rate of a given unit depends on the task variables in **b-e**. The position of each unit within a square is arbitrary; we therefore sorted them according to the amplitude of the coefficient of choice, which accounts for the diagonal bands of color in **b** (*top-left* to *bottom-right*, high to low choice coefficient). The positions of the pixels established in **b** are maintained in **c-e**, so that one can compare the amplitude of the coefficient for each task variable for every unit recorded from monkey A. Each of the four panels can be interpreted as the pattern of population activity elicited by the corresponding task variable. The four task variables elicit very distinct patterns of activity and are thus separable at the level of the population. Importantly, the coefficients were de-noised with principal component analysis (see Suppl. Information, section 6.7) and can be estimated reliably from noisy neural responses (Extended Data Fig. 4i-l). Differences between activation patterns therefore reflect differences in the properties of the underlying units, not noise. **f-j**, Recording locations and task-related patterns of population activity for monkey F. Same conventions as

in **a-e**. Recordings (**f**) covered the entire depth of the AS. The patterns of population activity elicited by a choice (**g**), by the motion evidence (**h**), and by context (**j**) are distinct, meaning that the representations of these task variables are separable at the level of the population. The representations of choice (**g**) and color (**i**), however, are not separable in monkey F, suggesting that color inputs are processed differently in the two monkeys (see main text).



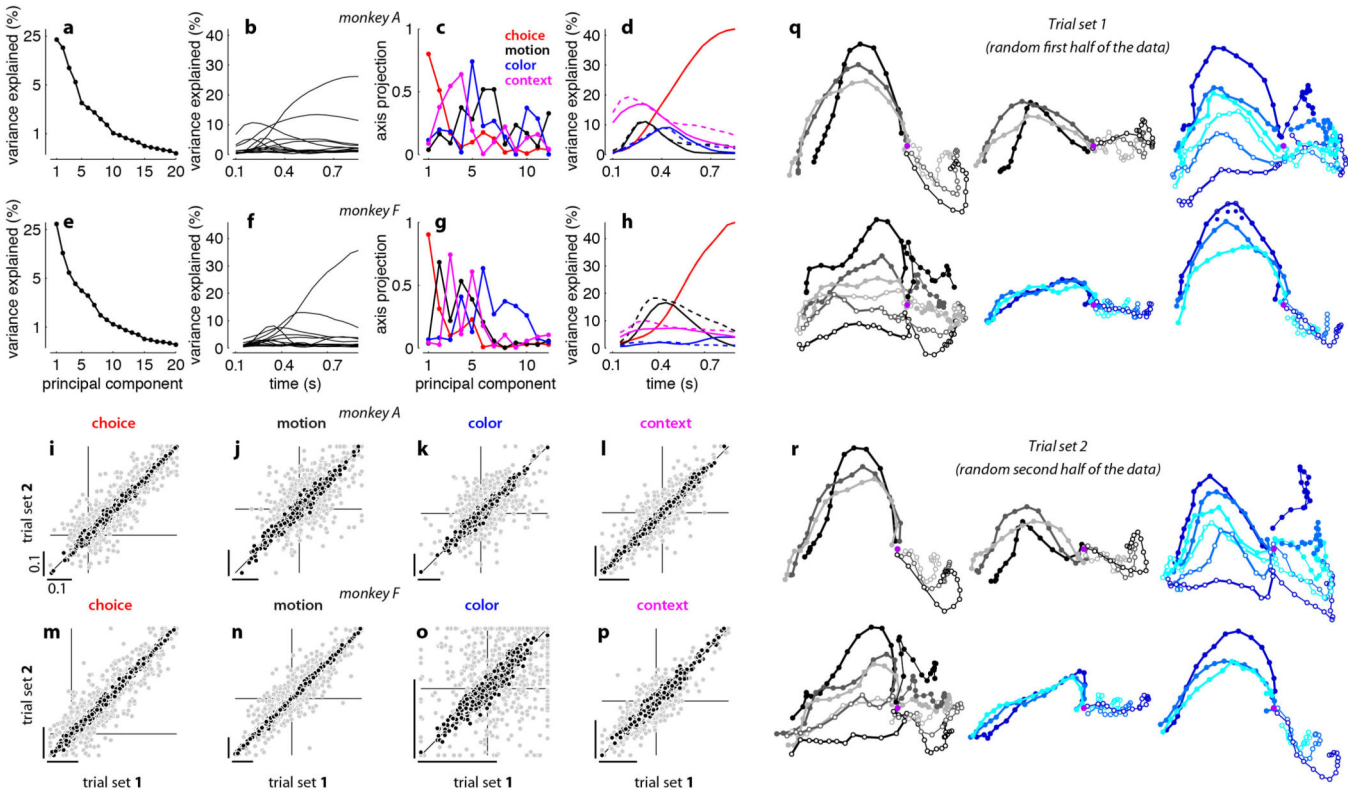
Extended Data Figure 2. Psychophysical performance for monkey F and for the model

a-d, Psychophysical performance for monkey F, for motion (*top*) and color contexts (*bottom*), averaged over 60 recording sessions (123,550 trials). Performance is shown as a function of motion (*left*) or color (*right*) coherence in each behavioral context. As in Fig. 1c-f, coherence values along the horizontal axis correspond to the average low, intermediate and high motion coherence (**a,c**) and color coherence (**b,d**) computed over all behavioral trials. The curves are fits of a behavioral model (see Suppl. Information, section 4). **e-h**, 'Psychophysical' performance for the trained neural-network model (Figs. 4–6) averaged over a total of 14,400 trials (200 repetitions per condition). Choices were generated based on the output of the model at the end of the stimulus presentation—an output larger than zero corresponds to a choice to the left target (choice 1), and an output smaller than zero corresponds to a choice to the left target (choice 2). We simulated model responses to inputs with motion and color coherences of 0.03, 0.12, and 0.50. The variability in the input (i.e. the variance of the underlying Gaussian distribution) was chosen such that the performance of the model for the relevant sensory signal qualitatively matches the performance of the monkeys. As in Fig. 1c-f, performance is shown as a function of motion (*left*) or color (*right*) coherence in the motion (*top*) and color contexts (*bottom*). Curves are fits of a behavioral model (as in **a-d** and in Fig. 1c-f). In each behavioral context, the relevant sensory input affects the model's choices (**e, h**), but the irrelevant input does not (**f, g**), reflecting successful context-dependent integration. In fact, the model output essentially corresponds to the bounded temporal integral of the relevant input (not shown) and is completely unaffected by the irrelevant input.



Extended Data Figure 3. Mixed representation of task variables in PFC

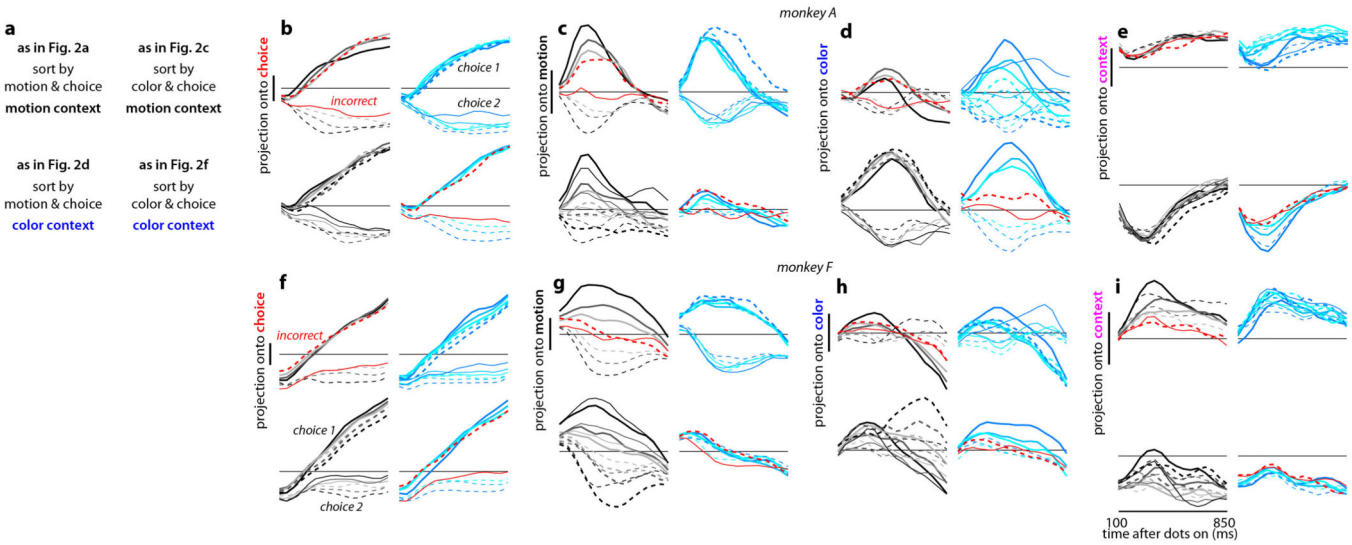
a-d, Example responses from 6 well-isolated single-units in monkey A. Each column shows average normalized responses on correct trials for one of the single-units. Responses are aligned to the onset of the random-dot stimulus, averaged with a 50ms sliding window, and sorted by one or more task-related variables (choice, motion coherence, color coherence, context). The green lines mark time intervals with significant effects of choice (**a**), motion coherence (**b**), color coherence (**c**), or context (**d**) as assessed by multi-variable, linear regression (regression coefficient different from zero, $p<0.05$). Linear regression and coefficient significance are computed over all trials (correct and incorrect, motion and color context; Supp. Information, section 6.3). The horizontal gray line corresponds to a normalized response equal to zero. **a**, Responses sorted by choice (*solid*: choice 1; *dashed*: choice 2) averaged over both contexts. **b**, Responses during motion context, sorted by choice and motion coherence (*black* to *light-gray*: high to low motion coherence). **c**, Responses during color context, sorted by choice and color coherence (*blue* to *cyan*: high to low color coherence). **d**, Responses sorted by choice and context (*black*: motion context; *blue*: color context). As is typical for PFC, the activity of the example units depends on many task variables, suggesting that they represent mixtures of the underlying task variables. **e-f**, De-noised regression coefficients for all units in monkey A (**e**) and monkey F (**f**). The data in Extended Data Fig. 1 are re-plotted here to directly compare the effects of different task variables (choice, motion, color, context) to each other. Each data point corresponds to a unit, and the position along the horizontal and vertical axes is the de-noised regression coefficient for the corresponding task variable. The horizontal and vertical lines in each panel intersect at the origin (0,0). Scale bars span the same range (0.1) in each panel. The different task variables are mixed at the level of individual units. While units modulated by only one of the task variables do occur in the population, they do not form distinct clusters but rather are part of a continuum that typically includes all possible combinations of selectivities. Significant correlations between coefficients are shown in *red* ($p<0.05$, Pearson's correlation coefficient r).



Extended Data Figure 4. Targeted dimensionality reduction of population responses, and reliability of task-related axes and population trajectories

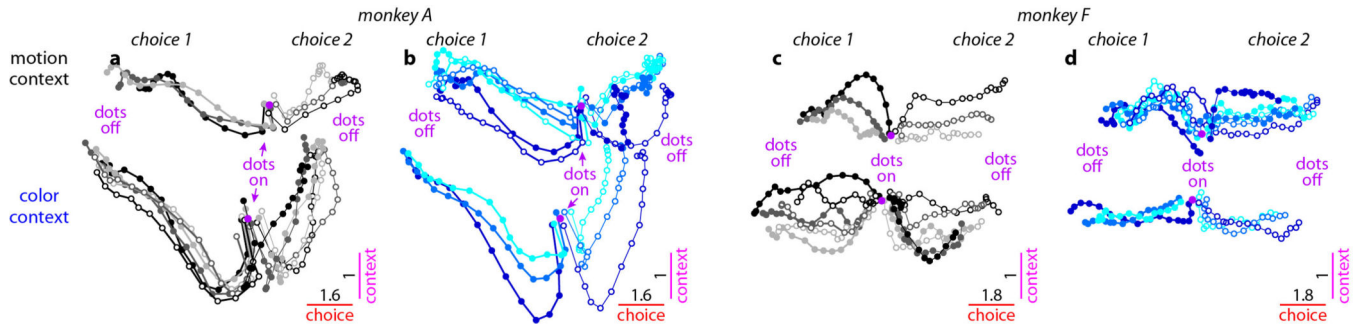
a, Fraction of variance explained by the first 20 principal components (PCs) of the responses in monkey A. PCs are computed on correct trials only, on condition-averaged responses. Conditions are defined based on choice, motion coherence, color coherence, and context. Each time point of the average response for a given condition contributes an 'independent' sample for the PC analysis, and variance is computed over conditions and times. **b**, Fraction of variance explained by the first 12 PCs. The total explainable variance (100%) is computed separately at each time, and reflects response differences across conditions. **c**, The four 'task-related axes' of choice, motion, color, and context expressed as linear combinations of the first 12 PCs. The four axes span a subspace containing the task-related variance in the population response (e.g. Fig. 2 and Extended Data Fig. 6) and are obtained by orthogonalizing the de-noised regression vectors for the corresponding task variables (see Suppl. information, section 6.7; de-noised regression coefficients are shown in Extended Data Figs. 1 and 3 **e,f**). The vertical axis in **c** corresponds to the projection of each axis onto a given PC (i.e. the contribution of that PC to each axis). All four axes project onto multiple PCs and thus the corresponding task variables are mixed at the level of single PCs. **d**, Fraction of variance explained by the task-related axes of choice, motion, color, and context (*solid lines*), as in **b**. The 4 axes explain a larger fraction of the variance than the PCs at many times but, unlike the PCs, they do not explain the variance common to all conditions that is due to the passage of time (not shown). A possible concern with our analysis is that the time courses of variance explained in **d** could be misleading if the task-related axes, which we estimated only at a single time for each variable, are in fact changing over time

during the presentation of the random dots. Under this scenario, for example, the “humped” shape of the motion input (*solid black trace*) might reflect a changing ensemble code for motion rather than actual changes in the strength of the motion signal in the neural population. To control for this possibility, we also computed time-varying ‘task-related axes’ by estimating the axes of motion, color and context separately at each time throughout the 750ms dots presentation. The fractions of variance explained by the time-varying axes (*dashed lines*) and by the fixed axes (*solid lines*) have similar amplitudes and time courses. Thus the effects of the corresponding task variables (during the presentation of the random dots) are adequately captured by the subspace spanned by the fixed axes (see Suppl. Information, section 6.8). **e-h**, Same as **a-d**, for monkey F. As shown in Extended Data Figs. 1g,i and 3f (*top-right* panel) the de-noised regression coefficients of color and choice are strongly correlated. As a consequence, the axis of color explains only a small fraction of the variance in the population responses (**h**, *blue*; see main text). **i-l**, Reliability of task-related axes in monkey A. To determine to what extent variability (i.e., noise) in single unit responses affects the task-related axes of choice, motion, color, and context (e.g. Fig. 2 and Extended Data Fig. 6), we estimated each axis twice from two separate sets of trials (trial sets 1 and 2 in **i-l**). For each unit, we first assigned each trial to one of two subsets, and estimated de-noised regression coefficients for the task variables separately for the two subsets. We then obtained task-related axes by orthogonalizing the corresponding de-noised coefficients (see Suppl. Information, section 6.9). Here, the orthogonalized coefficients are computed both with (*black*) and without (*gray*) PCA based de-noising. The horizontal and vertical lines in each panel intersect at the origin (0,0). Scale bars span the same range (0.1) in each panel. Data points lying outside the specified horizontal or vertical plotting ranges are shown on the corresponding edges in each panel. **i**, Coefficients of choice. Each data point corresponds to the orthogonalized coefficient of choice for a given unit, computed from trials in set 1 (*horizontal axis*) or in set 2 (*vertical axis*). **j-l**, same as **i** for the orthogonalized coefficients of motion (**j**), color (**k**), and context (**l**). **m-p**, Orthogonalized regression coefficients for monkey F, as in **i-l**. Overall, after de-noising the orthogonalized coefficients are highly consistent across the two sets of trials. Therefore, the observed differences in the activation pattern elicited by different task variables (Extended Data Fig. 1) are not due to the noisiness of neural responses, but rather reflect differences in the properties of the underlying units. **q-r**, Reliability of population trajectories. To assess the reliability of the trajectories in Fig. 2, we estimated the task-related axes and the resulting population trajectories (same conventions as Fig. 2) twice from two separate sets of trials (as **i-l**, see Suppl. Information, section 6.9). As in the example trajectories shown in **q** (trial set 1) and **r** (trial set 2), we consistently obtained very similar trajectories across the two sets of trials. To quantify the similarity between the trajectories from the two sets, we used trajectories obtained from one set to predict the trajectories obtained from the other set (see Suppl. Information, section 6.9). On average across 20 randomly defined pairs of trial sets, in both monkeys the population responses from one set explain 94% of the total variance in the responses of the other set (95% for the example in **q** and **r**). These numbers provide a lower bound on the true reliability of trajectories in Fig. 2, with are based on twice as many trials as those in **q** and **r**.



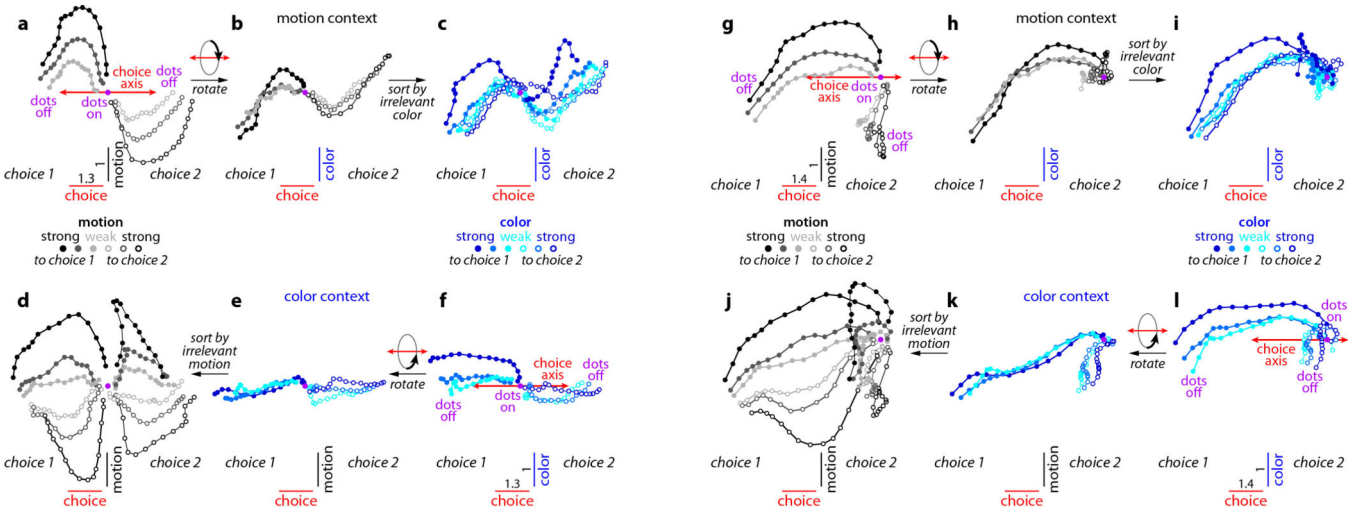
Extended Data Figure 5. Population responses along individual task-related axes

a-e, Responses for monkey A. The average population responses on correct trials are re-plotted from Fig. 2, together with responses on a subset of incorrect trials (*red curves*). Here the responses are represented explicitly as a function of time (*horizontal axis*) and projected separately (*vertical axes*) onto the axes of choice (**b**), motion (**c**), color (**d**), and context (**e**). As in Fig. 2, correct trials are sorted based on context (*motion*: top sub-panels; color: bottom sub-panels; see key in **a**), on the direction of the sensory evidence (*filled*: towards choice 1; *dashed*: towards choice 2) and strength of the sensory evidence (*black* to *light-gray*: strongest to weakest motion; *blue* to *cyan*: strongest to weakest color), and based on choice (*thick*: choice 1; *thin*: choice 2). Incorrect trials (*red curves*) are shown for the lowest motion coherence (during motion context, *top-left* in **b-e**) and the lowest color coherence (during color context, *bottom-right* in **b-e**). Vertical scale bars correspond to 1 unit of normalized response, and the horizontal lines are drawn at the same level in all four sub-panels within **b-e**. **a**, Key to the condition-averages shown in each panel of **b-e**, as well as to the corresponding state-space panels in Fig. 2. **b**, Projections of the population response onto the choice axis. Responses along the choice axis represent integration of evidence in both contexts. **c**, Projection onto the motion axis. Responses along the motion axis represent the momentary motion evidence during both motion (*top-left*) and color contexts (*bottom-left*) (curves are parametrically ordered based on motion strength in both contexts), but not the color evidence (*right*, curves are *not* ordered based on color strength). **d**, Projection onto the color axis. Responses along the color axis represent the momentary color evidence in the motion (*top-right*) and color contexts (*bottom-right*) (ordered), but not the motion evidence (*left*, not ordered). **e**, Projection onto the context axis. Responses in the motion context (*top*, all curves above the horizontal line) and color context (*bottom*, all curves below the horizontal line) are separated along the context axis, which maintains a representation of context. **f-i**, Responses for monkey F, same conventions as in **b-e**. The responses in **f-i** are also shown as trajectories in Extended Data Fig. 7g-l. The drift along the choice axis in Extended Data Fig. 7g-l is reflected in the overall positive slopes in **f**.



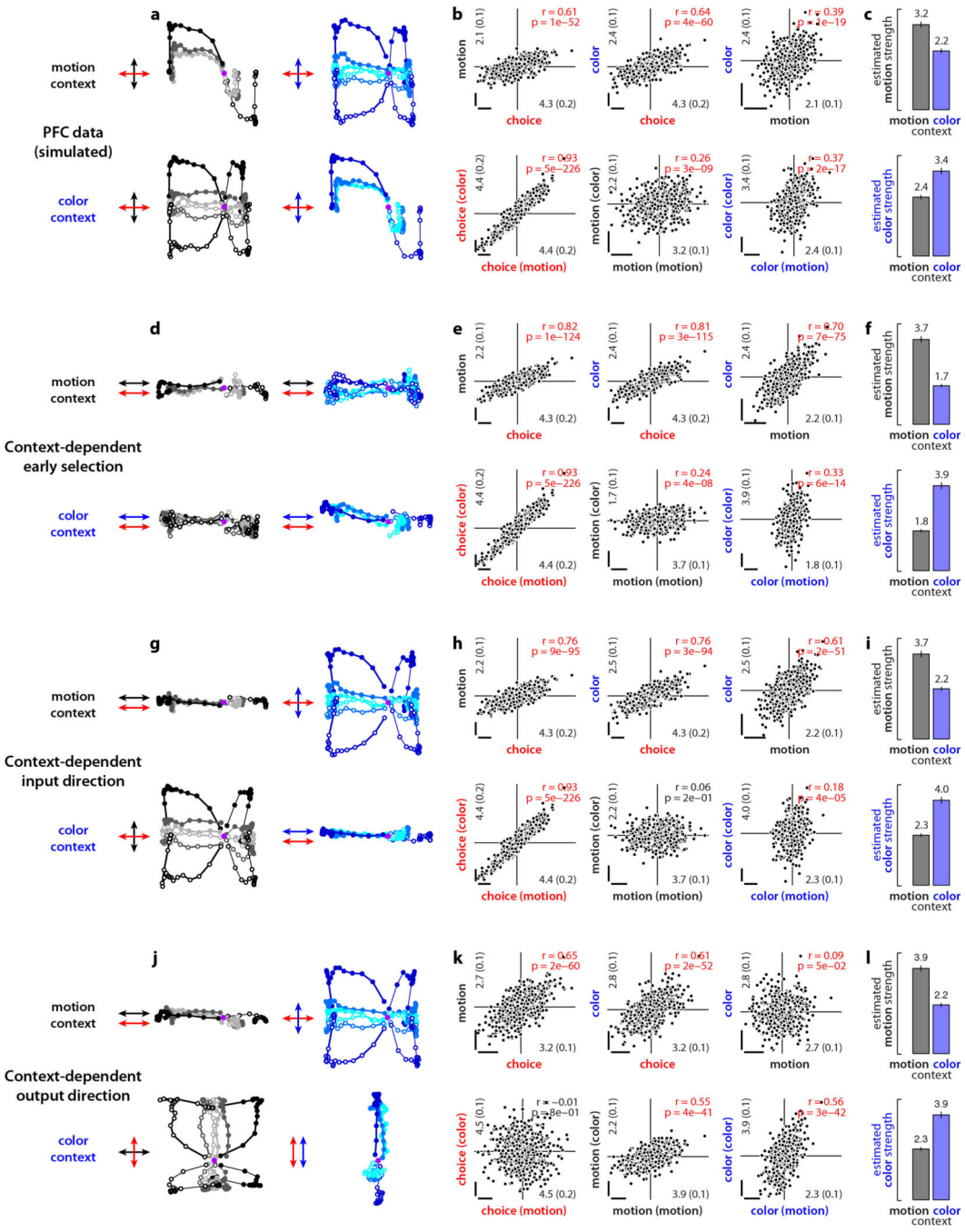
Extended Data Figure 6. Effect of context on PFC dynamics

a-b, Responses from monkey A. Same conditions and conventions as in Fig. 2, but for activity projected into the two-dimensional subspace capturing the variance due to choice (along the *choice* axis) and context (*context* axis). Components along the choice axis are enhanced relative to the context axis (see scale bars). The population response contains a representation of context, which is reflected in the separation between trajectories in the motion and color contexts along the axis of context. The contextual signal is strongest early during the dots presentation. **a**, Effects of context (*motion context* vs. *color context*), choice (*choice 1* vs. *choice 2*), and motion input (direction and coherence, *gray colors*). **b**, Same trials as in **a**, but averaged to show the effect of the color input (*blue colors*). **c-d**, Responses from monkey F, same conventions as in **a-b**. As in Extended Data Fig. 7a-f, we subtracted the across-condition average trajectory from each individual, raw trajectory (see Suppl. Information, section 6.10). The underlying raw population responses are shown in Extended Data Fig. 5f-i, and confirm that the representation of context is stable throughout the dots presentation time (Extended Data Fig. 5i).



Extended Data Figure 7. Dynamics of population responses in monkey F

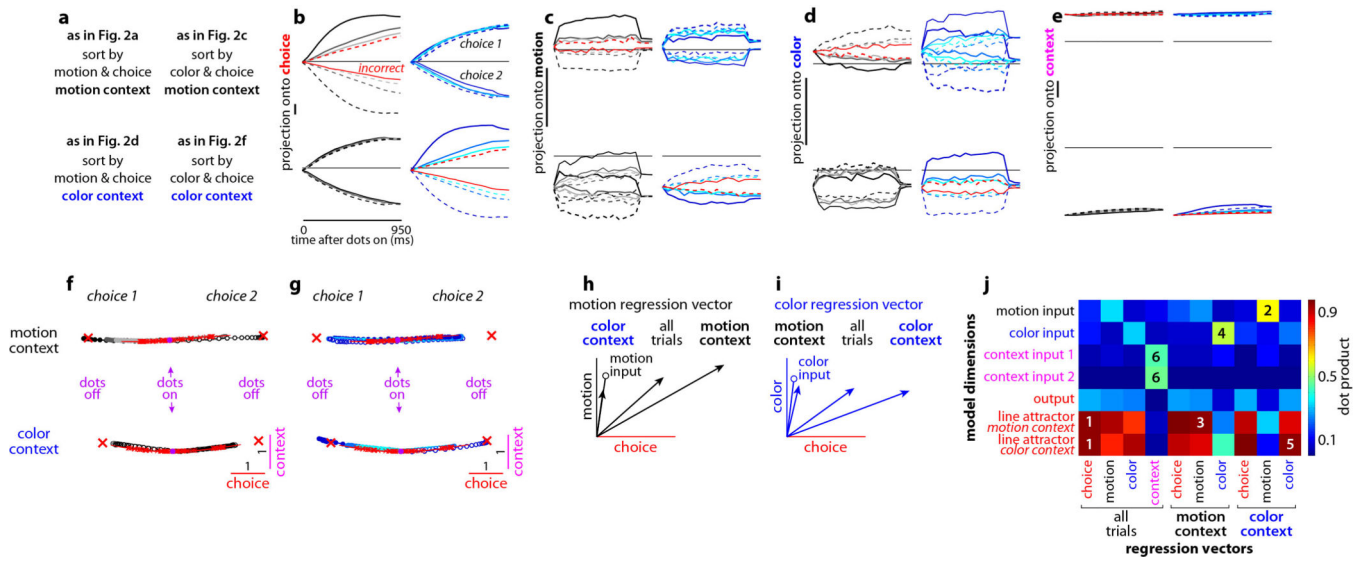
a-f, Response trajectories in the subspace spanned by the task related axes of choice, motion, and color. Same conventions as in Fig. 2. Unlike in Fig. 2, here we subtracted the across-condition average trajectory from each individual, raw trajectory (see Suppl. Information, section 6.10). The raw trajectories are shown in panels **g-l** and the corresponding projections onto individual axes in Extended Data Fig. 5f-i. Three key features of the population responses are shared in monkey A (Fig. 2) and monkey F. First, movement along a single choice axis (**a** and **f**, *red arrows*) corresponds to integration of the relevant evidence in both contexts. Second, in both contexts the momentary motion evidence elicits responses along the axis of motion, which is substantially different from the axis of choice (**a** and **d**). Third, the motion evidence is strongly represented whether it is relevant (**a**) or irrelevant (**d**). Thus, the processing of motion inputs in both monkeys is inconsistent with current models of selection and integration (Fig. 3b-d). Unlike in monkey A, responses along the color axis in monkey F (**f** and **c**) reflect the momentary color evidence only weakly. The effects of color on the trajectories in monkey F resemble the responses expected by the early selection model (Fig. 3b). **g-l**, Raw population responses. Population trajectories were computed and are represented as in Fig. 2. The trajectories in **a-f** were obtained by subtracting the across-condition average from each individual trajectory shown above. Overall, the responses have a tendency to move towards the left along the choice axis. An analogous, though weaker, overall drift can also be observed in monkey A, and contributes to the asymmetry between trajectories on choice 1 and choice 2 trials (Fig. 2). Because choice 1 corresponds to the target in the response field of the recorded neurons (see Suppl. Information, section 6.2), the drift reflects a tendency of individual firing rates to increase throughout the stimulus presentation time. By the definition of choice 1 and choice 2, a similar but opposite drift has to occur in neurons whose response field overlaps with choice 2 (whose responses we did not record). In the framework of diffusion-to-bound models, such a drift can be interpreted as an urgency signal, which guarantees that the decision boundary is reached before the offset of the dots (Reddi and Carpenter, 2000; Churchland, Kiani and Shadlen, 2008).



Extended Data Figure 8. Simulations of models of selective integration inconsistent with PFC responses

We simulated population responses mimicking the observed PFC responses (**a-c**) and alternative responses expected based on the three models of context-dependent selection described in Fig. 3b-d (**d-l**) (see Suppl. Information, section 8). These simulations are based on a diffusion-to-bound model, unlike the simulations of the recurrent neural network models in Figs. 5,6 and in Extended Data Figs. 9 and 10e-s. Here, single neurons represent mixtures of three time-dependent task variables of a diffusion-to-bound model, namely the

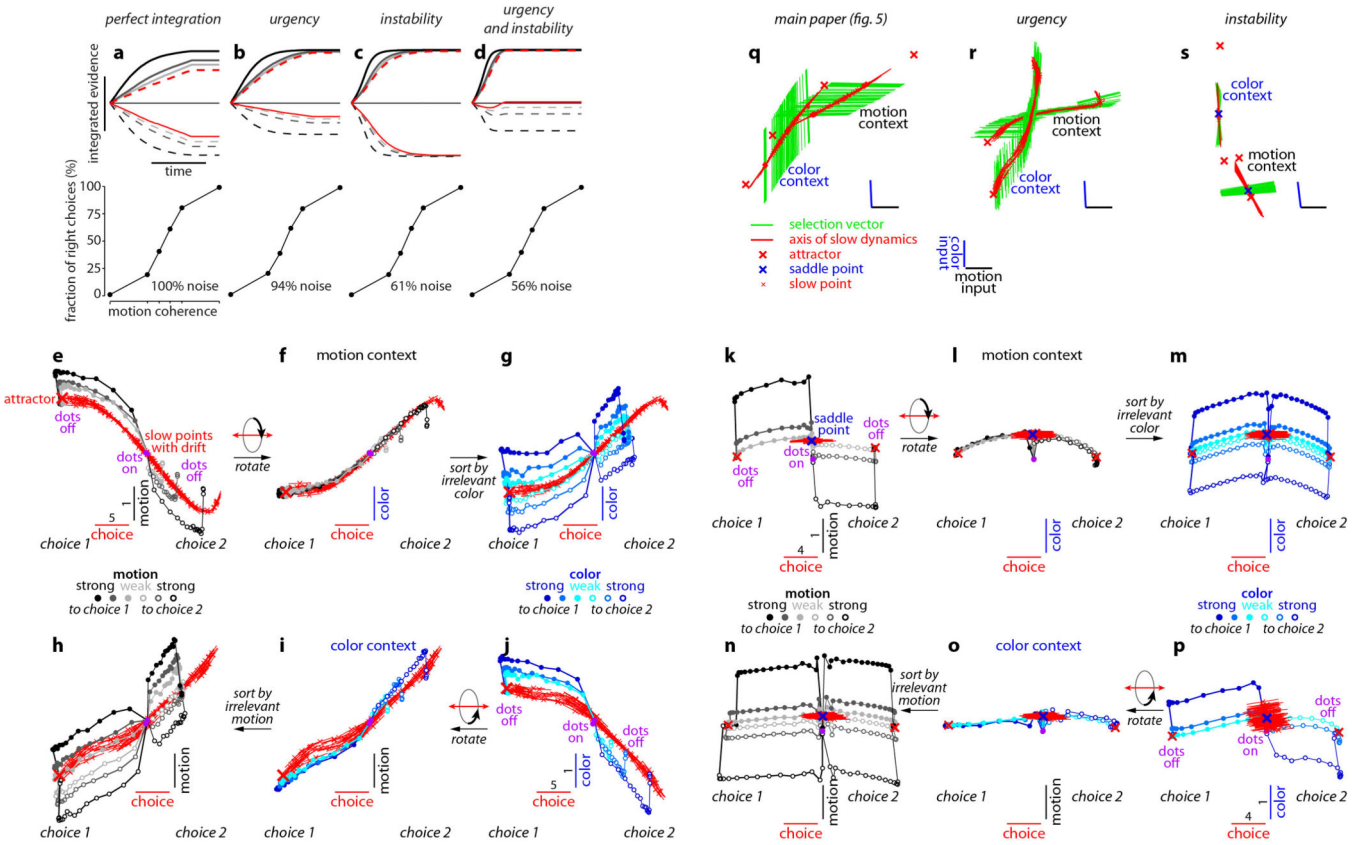
momentary motion and color evidence and the integrated relevant evidence. At the level of the population, these three task variables are represented along specific directions in state space (*arrows* in **a,d,g,j**; *red*, integrated evidence; *black*, momentary motion evidence; *blue*, momentary color evidence). The four simulations differ only with respect to the direction and context-dependence of the three task variables. We computed state space trajectories from the population responses using the targeted dimensionality reduction techniques discussed in the main text and in the Suppl. Information. The resulting simulated population responses reproduce the schematic population responses in Fig. 3. **a-c**, Simulated population responses mimicking the observed PFC responses (Fig. 2). **a**, Response trajectories in the 2d-subspace capturing the effects of choice and motion (*left*) or choice and color (*right*) in the motion (*top*) and color (*bottom*) contexts. Same conditions and conventions as in Fig 2a,c and Fig. 2d,f. The three task variables are represented along three orthogonal directions in state space (*arrows*). **b**, Regression coefficients of choice, motion, and color for all simulated units in the population. For each unit, coefficients were computed with linear regression on all simulated trials (*top*) or separately on trials from the motion or color context (*bottom*, context in parentheses). Scale bars represent arbitrary units. Numbers in the inset along each axis represent averages of the absolute value of the corresponding coefficients (±sem, in parentheses). **c**, Estimated strengths of the motion (*top*) and color (*bottom*) inputs during motion (*black*) and color (*blue*) contexts. Input strength is defined as the average of the absolute value of the corresponding regression coefficients. **d-f**, same as **a-c**, for simulated population responses expected from context-dependent early selection (Fig. 3b). When relevant, momentary motion (*top*) and color (*bottom*) evidence are represented along the same direction as integrated evidence (*arrows* in **d**). **g-i**, same as **a-c**, for simulated population responses expected from context-dependent input directions (Fig. 3c). Integrated evidence is represented along the same direction in both contexts (*red arrows* in **g**). The relevant momentary evidence (motion in the motion context, *top*; color in the color context, *bottom*) is aligned with the direction of integration, while the irrelevant momentary evidence is orthogonal to it (*black* and *blue* arrows in **g**). **j-l**, same as **a-c**, for simulated population responses expected from context-dependent output directions (Fig. 3d). The momentary motion and color evidence are represented along the same directions in both contexts (*black* and *blue arrows* in **j**). The direction of integration (*red arrows* in **j**) is aligned with the motion evidence in the motion context (*top*), and with the color evidence in the color context (*bottom*).



Extended Data Figure 9. Model population responses and validation of targeted dimensionality reduction

a-e, Model population responses along individual task-related axes, same conventions as in Extended Data Fig. 5. Here we defined the task-related axes directly based on the synaptic connectivity in the model (see Suppl. Information, section 7.6; and panels **h-j**), rather than using the approximate estimates based on the population response (as for the PFC data, e.g. Fig. 2). The same axes and the resulting projections underlie the trajectories in Fig. 5. The model integrates the contextually relevant evidence almost perfectly, and the responses along the choice axis (**b**) closely match the output of an appropriately tuned diffusion-to-bound model (not shown). Notably, near perfect integration is not a core feature of the proposed mechanism of context-dependent selection (see main text, and Extended Data Fig. 10). **f-g**, Effect of context on model dynamics, same conditions and conventions as in Extended Data Fig. 6. Network activity is projected onto the two-dimensional subspace capturing the variance due to choice (along the *choice* axis) and context (*context* axis). Same units on both axes (see scale bars). As in Fig. 5, fixed points of the dynamics (*red crosses*) and the associated right zero-eigenvectors (i.e. the local direction of the line attractor, *red lines*) were computed separately for motion (*top*) and color contexts (*bottom*) in the absence of sensory inputs. The line attractors computed in the two contexts, and the corresponding population trajectories, are separated along the context axis. **f**, Effects of context (*motion context, color context*), choice (*choice 1, choice 2*), and motion input (direction and coherence, *gray colors*) on the population trajectories. **g**, Same trials as in **f**, but re-sorted and averaged to show the effect of the color input (*blue colors*). The context axis is approximately orthogonal to the motion and color inputs, and thus the effects of motion and color on the population response (Fig. 5) are not revealed in the subspace spanned by the choice and context axes (**f** and **g**). **h-j**, Validation of targeted dimensionality reduction. To validate the dimensionality reduction approach used to analyze population responses in PFC (see Suppl. Information, sections 6.5–6.7), we estimated the regression vectors of choice, motion, color, and context from the *simulated* population responses (Fig. 5 and panels **b-g**) and compared them to the exactly known model dimensions that underlie the model dynamics (see definitions below). Here we estimated the regression vectors in three ways:

by pooling responses from all model units and all trials (as in the PFC data, e.g. Fig. 2 and Extended Data Fig. 6), or separately from the motion- and color-relevant trials (contexts). Orthogonalization of the regression vectors yields the task-related axes of the subspace of interest (e.g. axes in Fig. 2). Most model dimensions (motion, color, and context inputs, and output) were defined by the corresponding synaptic weights after training. The line attractor, on the other hand, is the average direction of the right zero-eigenvector of the linearized dynamics around a fixed point, and was computed separately for the motion and color contexts. **h**, The three regression vectors of motion (*black arrows*), plotted in the subspace spanned by the choice axis (i.e. the regression vector of choice) and the motion axis (i.e. the component of the regression vector of motion orthogonal to the choice axis). In the color context, the motion regression vector closely approximates the actual motion input (*black circle*—the model dimension defined by synaptic weights). During the motion context, however, the motion regression vector has a strong component along the choice axis, reflecting the integration of motion evidence along that axis. The motion regression vector estimated from all trials corresponds to the average of the vectors from the two contexts; thus all three motion regression vectors lie in the same plane. **i**, The three regression vectors of color (*blue arrows*) plotted in the subspace spanned by the choice and color axes, analogous to **h**. The color regression vector closely approximates the actual color input (*blue circle*) in the motion context, but has a strong component along the choice axis in the color context. Components along the motion (**h**) and color (**i**) axes are scaled by a factor of 2 relative to those along the choice axis. **j**, Dot products (*color bar*) between the regression vectors (*horizontal axis*) and the actual model dimensions (*vertical axis*), computed after setting all norms to 1. The choice regression vector closely approximates the direction of the line attractor in both contexts (squares labeled ‘1’). As shown also in **h** and **i**, the input regression vectors approximate the model inputs (defined by their synaptic weights) when the corresponding inputs are irrelevant (squares 2 and 4, motion and color), while they approximate the line attractor when relevant (squares 3 and 5). Thus the motion input is mostly contained in the plane spanned by the choice and motion axes (**h**), and the color input is mostly contained in the plane spanned by the choice and color axes (**i**). Finally, the single context regression vector is aligned with both context inputs (squares 6), and closely approximates the difference between the two (not shown).



Extended Data Figure 10. Urgency and instability in the integration process

a-d, Choice predictive neural activity (*top*) and psychometric curves (*bottom*) predicted by several variants of the standard diffusion-to-bound model (see Suppl. information, section 7.7). **a**, Standard diffusion-to-bound model. Noisy momentary evidence is integrated over time until one of two bounds (+1 or −1; choice 1 or choice 2) is reached. The momentary evidence at each time point is drawn from a Gaussian distribution whose mean corresponds to the coherence of the input, and whose fixed variance is adjusted in each model to achieve the same overall performance (i.e. similar psychometric curves, *bottom panels*). Coherences are 6, 18, and 50% (the average color coherences in monkey A, Fig. 1b). Average integrated evidence (neural firing rates, arbitrary units) is shown on choice 1 and choice 2 trials (*thick* vs. *thin*) for evidence pointing towards choice 1 or choice 2 (*solid* vs. *dashed*), on correct trials for all coherences (*light gray* to *black*, low to high coherence), and incorrect trials for the lowest coherence (*red*). The integrated evidence is analogous to the projection of the population response onto the choice axis (e.g. Extended Data Fig. 5b, *top-left* and *bottom-right*). **b**, Urgency model. Here the choice is determined by a race between two diffusion processes (typically corresponding to two hemispheres), one with bound at +1, the other with bound at −1. The diffusion in each process is subject to a constant drift towards the corresponding bound, in addition to the drift provided by the momentary evidence. The input-independent drift implements an ‘urgency’ signal, which guarantees that one of the bounds is reached within a short time. Only the integrated evidence from one of the diffusion processes is shown. The three ‘choice 1’ curves are compressed (in contrast to **a**) because the urgency signal causes the bound to be reached, and integration toward choice 1

to cease, more quickly than in **a**. In contrast, the 'choice 2' curves are not compressed since the diffusion process that accumulates evidence toward choice 1 never approaches a bound on these trials. **c**, Same as **a**, but here the diffusion process is subject to a drift away from the starting point (0) towards the closest bound (+1 or −1). The strength of the drift is proportional to the distance from the starting point, and creates an 'instability' at the starting point. **d**, Same as **b**, with an instability in the integration as in **c** for both diffusion processes. The asymmetry between choice 1 and choice 2 curves in **b** and **d** resembles the asymmetry in the corresponding PFC curves (Extended Data Figs. 5b,f, *upper left*). **e-j**, Neural network model with urgency. This model is based on a similar architecture as the model in Fig. 4. Unlike the neural network in Fig. 4, which was trained solely based on the model output on the last time bin of the trial, here the network is trained based on the output it produces throughout the entire input presentation. The network was trained to reproduce the integrated evidence (i.e. the decision variable) for one of the two diffusion processes (i.e. one of the two 'hemispheres') in a diffusion-to-bound model with urgency (**b**, see Suppl. Information, section 7.7). Similar conventions as in Fig. 5. The urgency signal is controlled by an additional binary input into the network. Here, the urgency and sensory inputs are turned off as soon as a bound is reached. The network generates only a single, stable fixed point in each context, corresponding to the decision boundary (*large red cross*). The model also implements a series of points of relatively slow dynamics (*small red crosses*) approximately lying on a single curve. The axes of slow dynamics at these slow points (*red lines*) are locally aligned. Notably, responses at these slow points have a strong tendency to drift towards the single, stable fixed point (the decision boundary), and thus the curve of slow points does not correspond to an approximate line attractor. This drift implements the urgency signal and causes an asymmetry in the trajectories, which converge on a single point for choice 1, but have endpoints that are parametrically ordered by coherence along the choice axis for choice 2. As discussed below (panel **r**), this model relies on the same mechanism of selection as the original model (Fig. 5, see main text). **k-p**, Neural network model with instability. Trajectories show simulated population responses for a model (same architecture as in Fig. 4) that was trained to solve the context-dependent task (Fig. 1) only on high-coherence stimuli and in the absence of internal noise (see Suppl. Information, section 7.7). Same conventions as in Fig. 5. In the absence of noise, prolonged integration of evidence is not necessary for accurate performance on the task. As a consequence, the model implements a saddle point (*blue cross*) instead of an approximate line attractor. Points of slow dynamics (*small red crosses*, obscured by the *red lines*) occur only close to the saddle point. The right-zero-eigenvectors of the linearized dynamics around these slow points (*red lines*) correspond to the directions of slowest dynamics, and determine the direction of the axis of choice. When displaced from the saddle point, the responses quickly drift towards one of the two stable attractors (*large red crosses*) corresponding to the choices. For a given choice, trajectories for all coherences therefore end in the same location along the choice axis, in contrast to the responses in the original model (Fig. 5). Despite these differences, the original model (Fig. 5) and the network model with instability (**k-p**) rely on a common mechanism of context-dependent selection (see panel **s**). **q-s**, Dynamical features (key, *bottom*) underlying input selection and choice in three related neural network models. All models are based on a common architecture (Fig. 4) but are the result of different training procedures. **q**, Dynamical features of the model described in the main paper (Figs. 5–6), re-

plotted from Fig. 6c. **r**, The urgency model (**e**-**j**). **s**, The instability model (**k**-**p**). In all models, the developing choice is implemented as more or less gradual movement along an axis of slow dynamics (specified by the locally computed right-eigenvectors associated with the near-zero eigenvalue of the linearized dynamics, *red lines*). The inputs are selected, i.e. result in movement along the axis of slow dynamics, depending on their projection onto the selection vector (the locally computed left-eigenvectors associated with the near-zero eigenvalue). In this sense, the three models implement the same mechanisms of context-dependent selection and choice.