

# Shared mechanisms underlie the control of working memory and attention

<https://doi.org/10.1038/s41586-021-03390-w>

Matthew F. Panichello<sup>1</sup> & Timothy J. Buschman<sup>1,2✉</sup>

Received: 15 April 2020

Accepted: 24 February 2021

Published online: 31 March 2021

 Check for updates

Cognitive control guides behaviour by controlling what, when, and how information is represented in the brain<sup>1</sup>. For example, attention controls sensory processing; top-down signals from prefrontal and parietal cortex strengthen the representation of task-relevant stimuli<sup>2–4</sup>. A similar ‘selection’ mechanism is thought to control the representations held ‘in mind’ – in working memory<sup>5–10</sup>. Here we show that shared neural mechanisms underlie the selection of items from working memory and attention to sensory stimuli. We trained rhesus monkeys to switch between two tasks, either selecting one item from a set of items held in working memory or attending to one stimulus from a set of visual stimuli. Neural recordings showed that similar representations in prefrontal cortex encoded the control of both selection and attention, suggesting that prefrontal cortex acts as a domain-general controller. By contrast, both attention and selection were represented independently in parietal and visual cortex. Both selection and attention facilitated behaviour by enhancing and transforming the representation of the selected memory or attended stimulus. Specifically, during the selection task, memory items were initially represented in independent subspaces of neural activity in prefrontal cortex. Selecting an item caused its representation to transform from its own subspace to a new subspace used to guide behaviour. A similar transformation occurred for attention. Our results suggest that prefrontal cortex controls cognition by dynamically transforming representations to control what and when cognitive computations are engaged.

To study the control of working memory and attention, we trained two monkeys to switch between two tasks. First, a retrospective (‘retro’) task required monkeys to select one of two items held in working memory (Fig. 1a). On each retro trial, the monkeys remembered the colours of two squares (colours drawn randomly from colour wheel) (Methods). After a memory delay, the monkeys were given a cue indicating whether to report the colour of the ‘upper’ or ‘lower’ square (now held in working memory). This cue was followed by a second memory delay, after which the monkeys reported the colour of the cued square by looking at the matching colour on a colour wheel (which was randomly rotated on each trial to prevent motor planning). Therefore, to perform the task, the monkeys held two colours in working memory, selected the colour of the cued square, and then used it to guide their response.

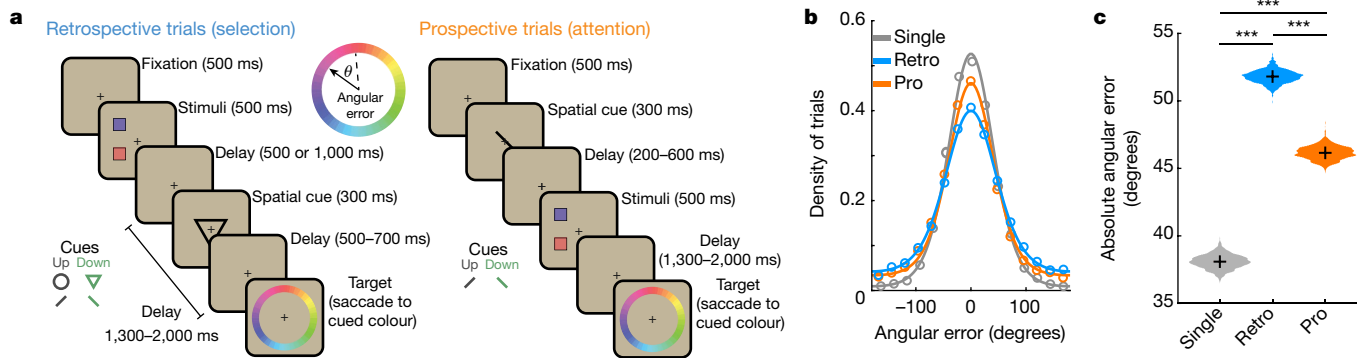
Monkeys performed the task well; the mean absolute angular error between the presented and reported colour was 51.8° (Fig. 1b, c, Extended Data Fig. 1a, b). As expected<sup>11–13</sup>, the error was reduced when only one item was presented (Fig. 1b, c, Extended Data Fig. 1d; the error was 38.1° for one item and 51.8° for two items ( $P < 0.001$ , randomization test)). The increased error with two items in memory is thought to be due to interference between the items<sup>14–17</sup>, which is reduced when an item is selected from working memory<sup>18,19</sup>. Consistent with this theory, the error was smaller when selection occurred earlier in the trial (Extended Data Fig. 1e, f; linear regression,  $\beta = 4.67^\circ \text{ s}^{-1} \pm 1.08$  (s.e.m.),  $P < 0.001$ , bootstrap).

In addition, monkeys performed a prospective (‘pro’) task. On pro trials, the cue was presented before the coloured squares, allowing the monkey to attend to the location of the to-be-reported stimulus (Fig. 1a). Consistent with attention reducing interference between stimuli<sup>20,21</sup> and modulating what enters working memory<sup>22</sup>, memory reports were more accurate in the pro task than the retro task (Fig. 1b, c, Extended Data Fig. 1d; 46.1° versus 51.8°;  $P < 0.001$ , randomization test) and increasing the number of stimuli from one to two led to a smaller increase in error on pro trials (9.01° versus 13.7° for pro versus retro;  $P < 0.001$ , bootstrap). These results highlight the functional homology between selection and attention, as both forms of control mitigate interference between representations<sup>14,20,23</sup>.

## Control of memory and attention

To understand the neural mechanisms of selection, and their relationship to attention, we simultaneously recorded from four regions involved in working memory and attention (Fig. 2a): lateral prefrontal cortex (LPFC; 682 neurons), frontal eye fields (FEF; 187 neurons), parietal cortex (Brodmann area 7a/b; 331 neurons), and intermediate visual area V4 (341 neurons). Consistent with previous work in humans<sup>9,24,25</sup>, neurons in all four regions carried information about which item was selected from working memory (that is, upper or lower) (Extended Data Fig. 2a, b). To quantify this information, we trained a logistic regression classifier to decode the location of selection from the firing rates of

<sup>1</sup>Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. <sup>2</sup>Department of Psychology, Princeton University, Princeton, NJ, USA. ✉e-mail: tbuschma@princeton.edu



**Fig. 1 | Monkeys use selection and attention to control the contents of working memory.** **a**, Time course of retro and pro tasks. Cues indicated whether the monkey should select the upper or lower item from working memory (retro task) or attend to the upper or lower item (pro task) and report that item after a delay. Reward was graded by error, calculated as the angular deviation between the cued and reported colour (dashed and solid lines in inset). On a subset of retro and pro trials, a single item was presented

(not shown). **b**, Distribution of error (circles) with best-fitting mixture models (lines) (Methods) for single item trials (grey), retro trials (blue), and pro trials (orange). As previously shown<sup>30</sup>, errors reflected both unsystematic error and systematic biases (Extended Data Fig. 1c). **c**, Bootstrapped distribution of mean absolute error in the retro, pro, and single-stimulus conditions ( $n = 8,620, 8,169$ , and  $4,207$  trials, respectively). \*\*\* $P < 0.001$ , two-sided uncorrected randomization test.

populations of neurons recorded in each region (Fig. 2b, Methods). The classifier found significant information about the location of selection in all four regions, emerging first in LPFC and then in posterior regions (Fig. 2c; 175 ms post-cue in LPFC, 245 ms in FEF, 285 ms in parietal, and 335 ms in V4). The emergence of information in LPFC was significantly earlier than in parietal and V4 ( $P = 0.005$  and  $P = 0.048$ , respectively; randomization test), but statistically indistinguishable from FEF ( $P = 0.371$ ). These results did not depend on the number of neurons recorded in each region and were not due to differences in neural responsiveness or noise (Extended Data Figs. 2, 3, Supplementary Table 1). Together, these results suggest that control of selection emerges first in prefrontal cortex and propagates to parietal and visual cortex.

Motivated by the functional homology between selection and attention<sup>5</sup>, we tested whether they were encoded in a shared population representation. Specifically, we tested whether the classifiers trained to decode the location of selection could generalize to decode the location of attention (and vice versa) (Fig. 2b, Methods). Consistent with a shared representation of selection and attention in LPFC, the ability of the classifiers to generalize in this way was significantly above chance and followed the time-course of the selection classifier (Fig. 2c). Individual LPFC neurons also generalized, representing the location of selection and attention similarly (Extended Data Fig. 4a–c;  $r(586) = 0.09, P = 0.036$ ).

By contrast, selection and attention were independently represented in FEF, parietal, and V4. Generalization was weaker in FEF and trended towards being delayed relative to LPFC (Fig. 2c;  $P = 0.12$ , randomization test). There was no significant generalization in parietal or V4 (Fig. 2c; this was not due to an inability to decode attention, Extended Data Fig. 4d, e). Consistent with different representations, the representations of selection and attention were uncorrelated in FEF, V4, and parietal neurons (Extended Data Fig. 4a–c; FEF:  $r(169) = 0.04, P = 0.617$ ; V4:  $r(318) = -0.04, P = 0.513$ ; parietal:  $r(301) = 0.03, P = 0.612$ ), although a positive correlation emerged later in FEF.

These results suggest that LPFC may act as a ‘domain-general’ controller, with a shared population representation that encodes both the selection of items from working memory and attention to sensory inputs. This could allow behaviours to generalize across working memory and sensory stimuli. By contrast, the task-specific representations seen in FEF and parietal (and partially in LPFC) could allow the specific control of memories or sensory stimuli. A combination of generalized and task-specific representations might balance the need to learn task-specific and generalized behaviours<sup>26,27</sup> (Supplementary Discussion 1).

### Selection and attention enhance memories

Next, we explored how selection and attention affected the neural representation of items in working memory. Single neurons in LPFC, FEF, parietal, and V4 all carried information about the colour of the upper or lower item (LPFC:  $n = 387$  of 607 cells; FEF: 114 of 178; parietal: 181 of 307; V4: 245 of 323; all  $P < 0.001$ , binomial test) (Methods, Extended Data Fig. 5a). In all four regions, information about the colour of the stimuli emerged during stimulus presentation and was maintained throughout the trial (Fig. 3, Extended Data Fig. 5b, Supplementary Discussion 2). These memory representations were related to behaviour: LPFC and V4 carried more information about the reported colour than the presented colour (Extended Data Fig. 6a;  $P < 0.001$ , randomization test).

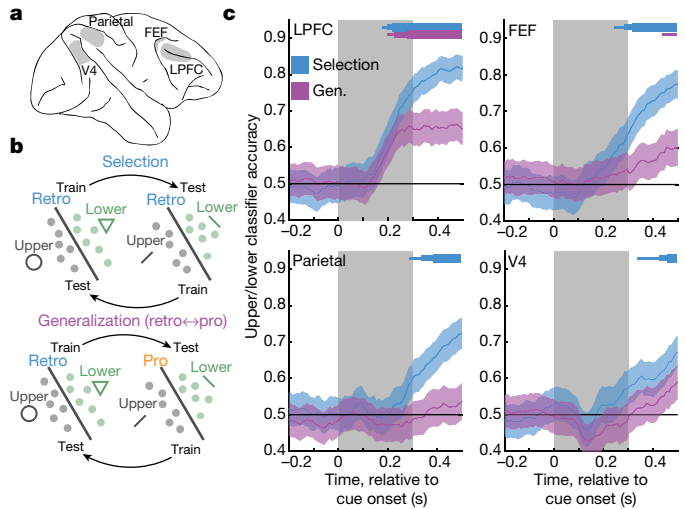
Consistent with previous work in humans<sup>6,8</sup>, selection strengthened memories in prefrontal and parietal cortex. In LPFC, colour information about the selected memory was greater than the information about the non-selected memory, starting 475 ms after cue onset (Fig. 3; also above pre-cue baseline, Extended Data Fig. 7a). Similar enhancements were seen in FEF and parietal (at 715 and 565 ms, respectively) (Fig. 3, Extended Data Fig. 7a).

The selective enhancement of a memory was related to behaviour in all four regions (Extended Data Fig. 7b, c). When memory reports were inaccurate, the effect of selection was absent in LPFC, FEF, and parietal. Although selection did not affect memory representations in V4 overall (Fig. 3b), information about the selected item was increased on trials with high memory accuracy and information about the non-selected item was increased on low-accuracy trials. These results suggest that memory errors occurred when the monkey failed to select an item or selected the wrong item.

Similar to selection, attention increased information about the attended stimuli, which suggests that similar mechanisms strengthen memory and sensory representations in prefrontal and parietal cortex (Extended Data Fig. 6b). However, in contrast to attention<sup>2,28</sup>, selection did not reduce information about the non-selected memory in LPFC and parietal (Extended Data Fig. 7a; but information did slightly decrease in FEF), which suggests that selection might not engage the competitive mechanisms that suppress unattended stimuli<sup>29</sup>.

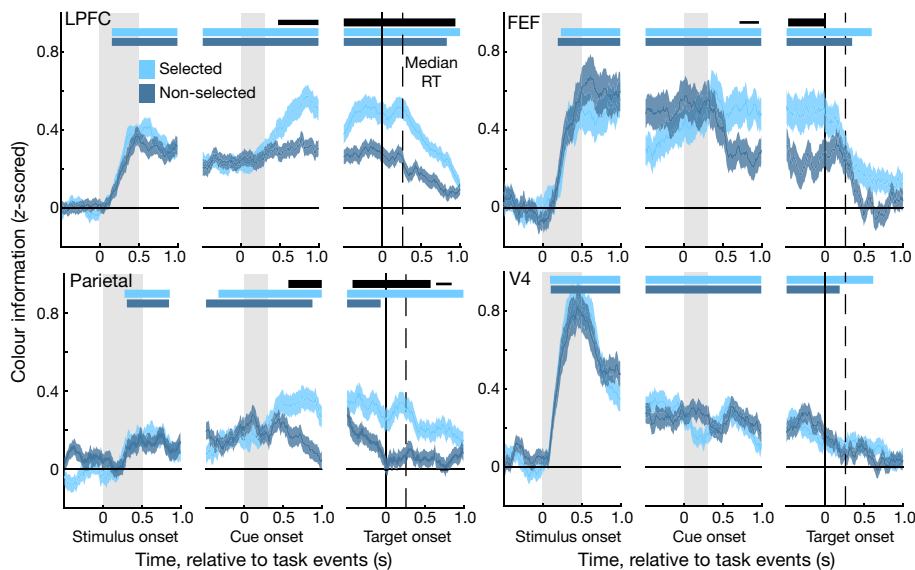
### Selection and attention transform memories

Finally, we were interested in how the changing task demands during retro trials affected memory representations. Early in the trial, before



**Fig. 2 | Selection is observed first in prefrontal cortex and shares a population code with attention.** **a**, Schematic of locations of neural recordings. **b**, Schematic of classifiers used to quantify information about whether the upper or lower item was selected from population firing rates (Methods). ‘Selection’ classifier accuracy was measured within retro trials (top) on held-out data. ‘Generalization’ classifier accuracy was measured across retro and pro trials (bottom). **c**, Time course of classifier accuracy for each brain region (labelled in top left). Lines and shading show mean  $\pm$  s.e.m. classification accuracy around cue onset for the selection (blue) and generalization classifiers (purple). Distribution reflects 1,000 iterations of classifiers, trained and tested on  $n = 60$  randomly sampled trials. Horizontal bars (top right of each plot) indicate above-chance classification ( $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick lines, respectively; one-sided uncorrected bootstrap).

selection, colour memories must be maintained in a form that allows the monkey to select the cued item (that is, colours are bound to a location). Later in the trial, after selection, only the colour of the selected item is needed to guide the visual search of the colour wheel and the



**Fig. 3 | Selection increases colour information in working memory.** Lines and shading show mean  $\pm$  s.e.m. z-scored colour information for the selected and non-selected colour (light and dark blue, respectively) in each brain region, averaged across neurons (LPFC,  $n = 570$ ; FEF,  $n = 163$ ; parietal,  $n = 292$ ; V4,  $n = 311$  neurons). Information was quantified by the circular entropy of each neuron’s response to colours (Methods). Horizontal bars at tops of plots indicate significant information for the selected (light blue) and non-selected item (dark

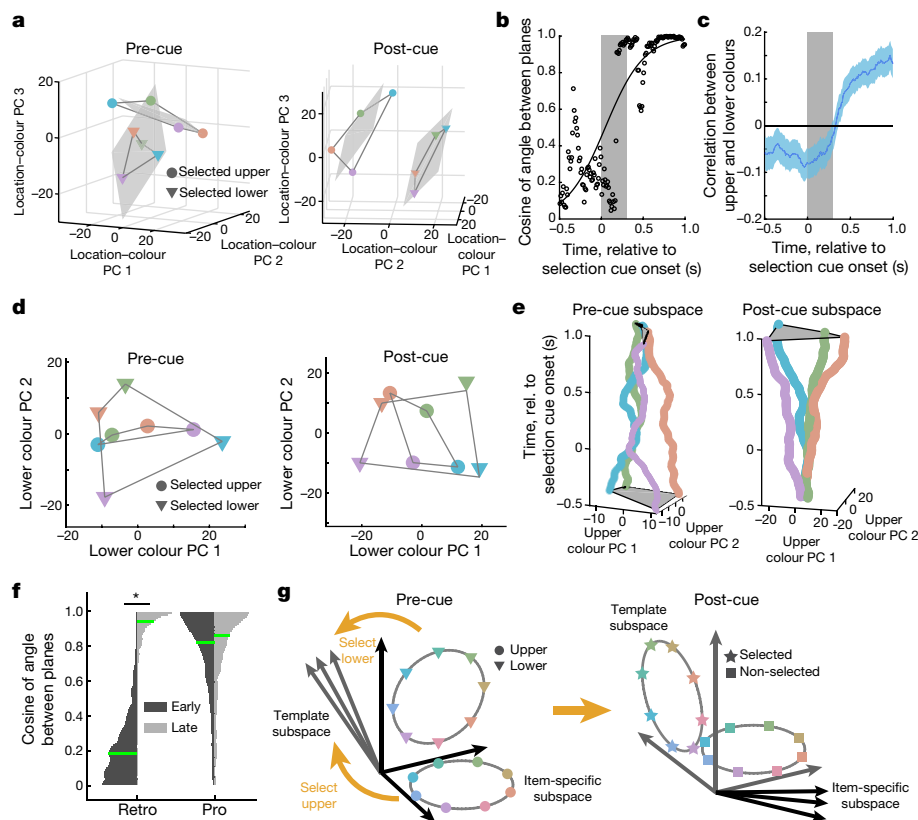
monkey’s response. Next, we show how selection transformed memory representations to match these changing task demands.

Before selection, the colour of each item in memory was represented in separate subspaces in the LPFC neural population. Figure 4a shows the representation of the colour of the upper and lower item, before selection (projected into a reduced 3D space) (Methods). Colour information showed a clear organization; the responses to four categories of colour were well separated and in colour order for both the upper and lower item (that is, neighbouring colours in colour space had neighbouring representations). Colour representations for each item were constrained to a ‘colour plane’, consistent with a 2D colour space (Methods). As seen in Fig. 4a, the upper and lower colour planes appeared to be independent from one another, suggesting that colour information about the upper and lower items was separated into two different item-specific subspaces in the LPFC population (before selection).

Consistent with the existence of separate subspaces, the median angle between the upper and lower colour planes was  $79.1^\circ$  (Fig. 4b; interquartile range (IQR),  $71.4\text{--}85.1^\circ$ ; Methods), which suggests that they were almost orthogonal before selection. This was not because the two items were encoded by separate populations of neurons. Rather, representations in LPFC overlapped<sup>27</sup>, with a significant proportion of neurons encoding both items (31% and 35% of neurons encoding the upper or lower item also encoded the other item;  $P = 1.21 \times 10^{-4}$ , binomial test) (Extended Data Fig. 8a, b). The colour planes were not completely orthogonal, as the representations of the upper and lower items were anti-correlated (Fig. 4c; for example, the  $N$ -neuron population vectors of ‘red upper’ and ‘red lower’ were anti-correlated; mean  $r = -0.067$  for  $-300$  to  $0$  ms pre-selection,  $P = 0.009$ , bootstrap). This modest anti-correlation might improve differentiation when the two items have similar colours.

Further supporting the existence of independent upper and lower subspaces before selection, colour representations of an item were less separated when they were projected onto the other subspace (Fig. 4d; each item’s subspace was defined as the 2D space that maximally captured colour information in the full  $N$ -dimensional neural space; Methods). To quantify the separability of colours, we measured the area of the quadrilateral defined by the four colour representations.

blue) and a significant difference in information about the selected and non-selected items (black). Bar thickness indicates significance:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick, respectively; two-sided cluster-corrected  $t$ -test. Stimulus colour information tended to emerge first in V4 (at 85 ms post-stimulus) and flow forward to LPFC (145 ms), FEF (185 ms), and then parietal (275 ms; V4 < parietal,  $P = 0.035$ ; FEF < parietal,  $P = 0.054$ ; randomization tests). By contrast, selection increased colour information in LPFC first (main text).



**Fig. 4 | Selection transforms memory information in a task-dependent manner.** **a**, Population response in LPFC for the colour of the selected item (binned into four colours indicated by marker colour; upper and lower indicated by marker shape). Population response is taken as the vector of the mean firing rate of neurons before the cue (pre-cue, left; 400 ms before cue) and after the cue (post-cue, right; before target onset). Responses are projected into a reduced dimensionality subspace defined by the first three principal components (PCs) of all eight colour–location pairs. Grey lines connect adjacent colours on colour wheel. Grey shaded regions show best fitting planes to each item. **b**, Cosine of the angle between the two colour planes (**a**) over time. Higher numbers reflect better alignment. Black line shows the best-fitting logistic function to  $n = 150$  time points. **c**, Mean correlation

between upper and lower colour representations in LPFC over time (line and shading show mean  $\pm$  s.e.m. over  $n = 1,000$  bootstrap resamples of trials). **d**, Colour representations in LPFC projected into the ‘lower’ subspace, before (left) and after (right) selection. Time points and markers as in **a**. **e**, ‘Upper’ colour representations in LPFC projected into the ‘upper’ subspace ( $x$ – $y$  axes) over time ( $z$ -axis, relative to selection). **f**, Histograms show distribution of the cosine of the angle between the best-fitting planes for the upper and lower stimuli in an ‘early’ (150–350 ms post-stimulus offset) and ‘late’ (200–0 ms before target onset) time period for retro (left) and pro (right) tasks ( $n = 1,000$  bootstrap resamples of trials). Green lines indicate median. Horizontal lines indicate pairwise comparisons.  $*P = 0.016$ , two-sided uncorrected bootstrap. **g**, Schematic of how selection transforms colour representations.

This ‘colour area’ was greater when colour representations were projected into their own subspace rather than the other subspace (reflecting greater separation in their own subspace; 86.1 versus 35.2 units<sup>2</sup>,  $P = 0.041$ , bootstrap; all subspaces defined on held-out data).

After selection, memory representations in LPFC were transformed into a different subspace (as previously theorized<sup>7</sup>). Reflecting this, the separation of colours in the pre-selection subspace collapsed by the end of the second memory delay (Fig. 4e (left), Extended Data Fig. 8c). Accordingly, colour area tended to decrease over time, from 74.1 to 39.4 units<sup>2</sup> in the pre-selection subspace (Extended Data Fig. 8d;  $P = 0.076$ , bootstrap). Instead, after selection, colours were represented in a new ‘post-selection’ subspace (Fig. 4e (right), Extended Data Fig. 8c, d; colour area in post-selection subspace increased from 27.8 to 261.9 units<sup>2</sup> over time,  $P < 0.001$ , bootstrap).

Whereas pre-selection subspaces were independent, the post-selection subspaces of the upper and lower items were aligned (Fig. 4a). The upper and lower colour planes were now parallel (angle between the planes was 20.1°; IQR, 11.6–29.0°). The cosine of the angle between the upper and lower colour planes increased after selection (Fig. 4b;  $P = 0.006$ , bootstrap test of logistic regression). Furthermore, the representations of the selected item’s colour shifted from being anti-correlated before selection to positively correlated after selection (Fig. 4c; mean  $r = 0.139$  for –300 to 0 ms before target

onset,  $P < 0.001$  versus zero and versus pre-cue, bootstrap). Finally, colour representations of an item were now well separated when they were projected onto the other colour subspace (Fig. 4d; colour area increased from 35.2 to 94.0 units<sup>2</sup> over time,  $P = 0.010$ , bootstrap). Together, these results suggest that selection transformed memories from independent item-specific subspaces to a common subspace that represented the colour of the selected item, regardless of its original location. Reflecting the importance of this transformation, the strength of alignment of colour spaces in LPFC was correlated with behaviour: when memory reports were inaccurate, the cosine of the angle between the two colour planes was reduced (Extended Data Fig. 9f,  $P = 0.027$ , randomization test).

The degree of transformation iteratively decreased in FEF, parietal, and V4 (Extended Data Fig. 9c, d). This decrease might reflect a gradient in the flexibility of neural responses across regions, with dynamic, integrative, representations in prefrontal cortex and more static, localized, representations in visual cortex.

Selection also transformed the non-selected memories in LPFC, although to a lesser degree: the colour planes of the non-selected items tended to become aligned (IQR, 61.4–83.7° to 15.5–39.7°;  $P = 0.085$ , bootstrap) but the post-cue representations were not significantly correlated ( $P = 0.202$  against zero) (Extended Data Fig. 9a–d). Critically, the non-selected item remained nearly orthogonal to the selected item

before and after selection (IQR: 80.1–85.5° to 75.7–82.4° for pre- and post-cue;  $P = 0.287$ , bootstrap) (Extended Data Fig. 9a, c, d), which could avoid interference between the selected and unselected item. Notably, the transformations acting on the selected and non-selected representations partially generalized to the other item, suggesting that the transformation had a common component that acted on both items simultaneously (Extended Data Fig. 9e, Supplementary Discussion 3).

As noted above, the dynamic re-alignment of neural representations reflects the changing task demands during the trial: independently encoding items before selection but aligning items after selection, abstracting over item location. Consistent with the transformation of memories being driven by task demands, memory representations were aligned immediately after stimulus presentation on pro trials. In LPFC, the representations of the upper and lower colours were positively correlated after stimulus offset on pro trials (Extended Data Fig. 10a–d). In addition, the upper and lower colour planes were well aligned throughout the trial (Fig. 4f, Extended Data Fig. 10e; early: median angle 34.5°, IQR 22.1–51.4°; late: median angle, 30.4°, IQR 18.5–46.2°; no change with time,  $P = 0.449$ ; there was a trend towards an interaction between pre/post and pro/retro,  $P = 0.067$ , bootstrap).

The same aligned subspace seemed to be used in retro and pro trials: there was a weak, but significant, correlation between colour representations at the end of the delay on pro and retro trials (Extended Data Fig. 10f, mean  $\rho = 0.06$ ,  $P = 0.015$ , bootstrap). This correlation did not exist before selection (mean  $\rho = -0.01$ ,  $P = 0.634$ ) and increased with time ( $P = 0.027$ , bootstrap).

The task-dependent dynamic transformations we have observed might allow the cognitive control of behaviour. In the retro task, selection transformed colour information from independent, item-specific subspaces to a shared 'template' subspace (Fig. 4g). From the perspective of a neural circuit decoding information from the template subspace to guide visual search, the transformation abstracts over location and allows the selected item to guide the monkey's response. As the item-specific and non-selected subspaces are orthogonal to the template subspace, this circuit would be unaffected by those representations. In this way, the timing of the transformation determines when this circuit is engaged (for example, after selection in the retro task or immediately in the pro task). Thus, cognitive control may dynamically transform representations to control what and when cognitive computations are engaged.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-021-03390-w>.

1. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
2. Buschman, T. J. & Kastner, S. From behavior to neural dynamics: an integrated theory of attention. *Neuron* **88**, 127–144 (2015).
3. Buschman, T. J. & Miller, E. K. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* **315**, 1860–1862 (2007).
4. Moore, T. & Armstrong, K. M. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* **421**, 370–373 (2003).
5. Gazzaley, A. & Nobre, A. C. Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* **16**, 129–135 (2012).
6. Sprague, T. C., Ester, E. F. & Serences, J. T. Restoring latent visual working memory representations in human cortex. *Neuron* **91**, 694–707 (2016).
7. Myers, N. E., Stokes, M. G. & Nobre, A. C. Prioritizing information during working memory: beyond sustained internal attention. *Trends Cogn. Sci.* **21**, 449–461 (2017).
8. Ester, E. F., Nouri, A. & Rodriguez, L. Retrospective cues mitigate information loss in human cortex during working memory storage. *J. Neurosci.* **38**, 8538–8548 (2018).
9. Nobre, A. C. et al. Orienting attention to locations in perceptual versus mental representations. *J. Cogn. Neurosci.* **16**, 363–373 (2004).
10. Murray, A. M., Nobre, A. C., Clark, I. A., Cravo, A. M. & Stokes, M. G. Attention restores discrete items to visual short-term memory. *Psychol. Sci.* **24**, 550–556 (2013).
11. Wilken, P. & Ma, W. J. A detection theory account of change detection. *J. Vis.* **4**, 1120–1135 (2004).
12. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
13. Bays, P. M., Catalao, R. F. G. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* **9**, 7 (2009).
14. Buschman, T. J., Siegel, M., Roy, J. E. & Miller, E. K. Neural substrates of cognitive capacity limitations. *Proc. Natl Acad. Sci. USA* **108**, 11252–11255 (2011).
15. Sprague, T. C., Ester, E. F. & Serences, J. T. Reconstructions of information in visual spatial working memory degrade with memory load. *Curr. Biol.* **24**, 2174–2180 (2014).
16. Bays, P. M. Spikes not slots: noise in neural populations limits working memory. *Trends Cogn. Sci.* **19**, 431–438 (2015).
17. Bouchacourt, F. & Buschman, T. J. A flexible model of working memory. *Neuron* **103**, 147–160.e8 (2019).
18. Pertzov, Y., Bays, P. M., Joseph, S. & Husain, M. Rapid forgetting prevented by retrospective attention cues. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 1224–1231 (2013).
19. Bays, P. M. & Taylor, R. A neural model of retrospective attention in visual working memory. *Cognit. Psychol.* **100**, 43–52 (2018).
20. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
21. Trueue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
22. Everling, S., Tinsley, C. J., Gaffan, D. & Duncan, J. Filtering of neural signals by focused attention in the monkey prefrontal cortex. *Nat. Neurosci.* **5**, 671–676 (2002).
23. Schneegans, S. & Bays, P. M. Restoration of fMRI decodability does not imply latent working memory states. *J. Cogn. Neurosci.* **29**, 1977–1994 (2017).
24. Nee, D. E. & Jonides, J. Common and distinct neural correlates of perceptual and memorial selection. *Neuroimage* **45**, 963–975 (2009).
25. Quentin, R. et al. Differential brain mechanisms of selection and maintenance of information during working memory. *J. Neurosci.* **39**, 3728–3740 (2019).
26. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020).
27. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
28. Reynolds, J. H., Chelazzi, L. & Desimone, R. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* **19**, 1736–1753 (1999).
29. Reynolds, J. H. & Heeger, D. J. The normalization model of attention. *Neuron* **61**, 168–185 (2009).
30. Panichello, M. F., DePasquale, B., Pillow, J. W. & Buschman, T. J. Error-correcting dynamics in visual working memory. *Nat. Commun.* **10**, 3366 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

## Methods

### Subjects

Two adult (8–9 years old) male rhesus macaques (*Macaca mulatta*) participated in the experiment. Monkeys 1 and 2 weighed 12.1 and 8.9 kg, respectively. All experimental procedures were approved by the Princeton University Institutional Animal Care and Use Committee and were in accordance with the policies and procedures of the National Institutes of Health.

The subject number was chosen to be consistent with previous work. Both monkeys performed the same experiments and so no randomization or blinding of monkey identity was necessary. As detailed below, conditions within each experiment were chosen randomly and experimenters were blind to experimental conditions when pre-processing the data.

### Behavioural task

Stimuli were presented on a Dell U2413 LCD monitor positioned at a viewing distance of 58 cm using Psychtoolbox and MATLAB (Mathworks). The monitor was calibrated using an X-Rite iDisplay Pro colorimeter to ensure accurate colour rendering. During the experiment, subjects were required to remember the colour of either 1 or 2 square stimuli presented at two possible locations. The colour of each sample was drawn randomly from 64 evenly spaced points along a photometrically isoluminant circle in CIELAB colour space. This circle was centred at ( $L = 60$ ,  $a = 6$ ,  $b = 14$ ) and the radius was 57 units. Colours were independent across locations. The stimuli measured  $2^\circ$  of visual angle (DVA) on each side. Each stimulus could appear at one of two possible spatial locations:  $45^\circ$  clockwise or anticlockwise from the horizontal meridian (in the right hemifield; stimuli are depicted in the left hemifield in Fig. 1 for ease of visualization) with an eccentricity of 5 DVA from fixation. To perform the retrospective task, the monkey had to remember which colour was at each location (that is, the 'upper' and 'lower' colours).

The monkeys initiated each trial by fixating a cross at the centre of the screen. On retro trials, after 500 ms of fixation, one (20% of trials) or two (80% of trials) stimuli appeared on the screen. The stimuli were displayed for 500 ms, followed by a memory delay of 500 or 1,000 ms. Next, a symbolic cue was presented at fixation for 300 ms. This cue indicated which sample (upper or lower) the monkey should report to get a juice reward. The location of the selected memory was randomly chosen on each trial. Two sets of cues were used in the experiment to dissociate the meaning of the cue from its physical form. The first set (cue set 1) consisted of lines oriented  $45^\circ$  clockwise and anticlockwise from the horizontal meridian (cueing the lower and upper stimulus, respectively). The second set (cue set 2) consisted of a triangle or a circle (cueing the lower or upper stimulus, respectively). Cues were presented at fixation and subtended 2 DVA. After the cue, there was a second memory delay (500–700 ms), after which a response screen appeared. The response screen consisted of a ring  $2^\circ$  thick with an outer radius of  $5^\circ$ . The monkeys made their response by breaking fixation and saccading to the section of the colour wheel that corresponded to the colour of the selected (cued) memory. In previous work using human subjects, observers are typically free to foveate the colour wheel and fine-tune their selection, so differences in performance between monkeys and humans<sup>30</sup> may in part reflect task design. The colour ring was randomly rotated on each trial to prevent motor planning or spatial encoding of memories. The monkeys received a graded juice reward that depended on the accuracy of their response. The number of drops of juice awarded for a response was determined according to a circular normal (von Mises) distribution centred at  $0^\circ$  error with a standard deviation of  $22^\circ$ . This distribution was scaled to have a peak amplitude of 12, and non-integer values were rounded up. When response error was greater than  $60^\circ$  for monkey 1 ( $40^\circ$  for monkey 2), no juice was awarded and the monkey experienced a short time-out of 1–2 s. Responses had to

be made within 8 s, although, in practice, this restriction was unnecessary as response times were on the order of 200–300 ms.

Pro trials were similar to retro trials, except that the cue was presented 200–600 ms before the stimuli. After the coloured squares, a single continuous delay occurred before the onset of the response screen (1,300–2,000 ms for monkey 1 and 1,000–2,000 ms for monkey 2). For behavioural analyses and all neural analyses around the response epoch, we analysed only trials with a minimum delay of 1,300 ms to match the total delay range for pro and retro trials.

Condition (retro or pro) and cue set were manipulated in a blocked fashion. Monkeys transitioned among three different block types: (1) pro trials using cue set 1, (2) retro trials using cue set 1, and (3) retro trials using cue set 2. The sequence of blocks was random. Transitions between blocks occurred after the monkey had performed 60 correct trials of block type 1 (pro) or 30 correct trials for block types 2 and 3 (retro), balancing the total number of pro and retro trials. All electrophysiological recordings were done during this task.

In addition, both monkeys completed a second behavioural experiment (experiment 2) without electrophysiological recordings (Extended Data Fig. 1e). In experiment 2, all trials were a variant of the retrospective load 2 condition, the total stimulus–target delay was fixed to 2,400 ms, and the stimulus–cue delay was randomly selected to be 500, 1,000, or 1,500 ms. This manipulation allowed us to test whether the timing of the retrocue affected the accuracy of memory, thereby isolating the effect of selection on the contents of working memory.

The eye position of the monkeys was continuously monitored at 1 kHz using an Eyelink 1000 Plus eye-tracking system (SR Research). The monkeys had to maintain their gaze within a  $2^\circ$  circle around the central cross during the entire trial until the response. If they did not maintain fixation, the trial was aborted, and the monkey received a brief timeout.

We analysed all completed trials, defined as any trial on which the monkey successfully maintained fixation and made a saccade to the colour wheel, regardless of accuracy. Monkey 1 completed 9,865 trials over 10 sessions and monkey 2 completed 11,131 trials over 13 sessions.

As shown in Extended Data Fig. 1, the behaviour of the two monkeys was qualitatively similar and so we pooled data across monkeys for all analyses.

### Surgical procedures and recordings

Monkeys were implanted with a titanium headpost to immobilize the head and with two titanium chambers for providing access to the brain. The chambers were positioned using 3D models of the brain and skull obtained from structural MRI scans. Chambers were placed to allow electrophysiological recording from LPFC, FEF, parietal, and V4.

Epoxy-coated tungsten electrodes (FHC) were used for both recording and microstimulation. Electrodes were lowered using a custom-built microdrive assembly that lowered electrodes in pairs from a single screw. Recordings were acute; up to 80 electrodes were lowered through the intact dura at the beginning of each recording session and allowed to settle for 2–3 h before recording. This enabled stable isolation of single units over the session. Broadband activity (sampling frequency, 30 kHz) was recorded from each electrode (Blackrock Microsystems). We performed 13 recording sessions with monkey 2 and 10 sessions with monkey 1.

After recordings were complete, we confirmed electrode locations by performing structural MRIs after lowering two electrodes in each chamber into the cortex. Using the shadows of these two electrodes, the positions of the other electrodes in each chamber could be reconstructed. Electrodes were categorized as falling into LPFC, FEF, parietal, or V4 based on anatomical landmarks.

In separate experiments, we identified which electrodes were located in FEF using electrical microstimulation. On the basis of previous work<sup>31</sup>, we defined FEF sites as those for which electrical stimulation elicited a saccadic eye movement. Electrical stimulation was delivered in 200-ms trains of anodal-leading biphasic pulses with a width of 400  $\mu$ s and

an inter-pulse frequency of 330 Hz. Electrical stimulation was delivered to each electrode in the frontal well of each monkey and FEF sites were identified as those sites for which electrical stimulation (<50  $\mu$ A) consistently evoked a saccade with a stereotyped eye movement vector at least 50% of the time. Untested electrode sites (for example, from recordings on days with a different offset in the spatial distribution of electrodes) were classified as belonging to FEF if they fell within 1 mm of confirmed stimulation sites and were positioned in the anterior bank of the arcuate sulcus (as confirmed via MRI).

### Signal preprocessing

Electrophysiological signals were filtered offline using a 4-pole 300 Hz high-pass Butterworth filter. For monkey 1, to reduce common noise, the voltage time series  $x$  recorded from each electrode was re-referenced to the common median reference<sup>32</sup> by subtracting the median voltage across all electrodes in the same recording chamber at each time point.

The spike detection threshold for all recordings was set equal to  $-4\sigma_n$ , in which  $\sigma_n$  is an estimate of the standard deviation of the noise distribution:

$$\sigma_n = \text{median}\left(\frac{|x|}{0.6745}\right)$$

Time points at which  $x$  crossed this threshold with a negative slope were identified as putative spiking events. Repeated threshold crossings within 32 samples (1.0667 ms) were excluded. Waveforms around each putative spike time were extracted and were manually sorted into single units, multi-unit activity, or noise using Plexon Offline Sorter (Plexon).

For all analyses, spike times of single units were converted into smoothed firing rates (sampling interval, 10 ms) by representing each spiking event as a delta function and convolving this time series with a causal half-Gaussian kernel ( $\sigma = 200$  ms).

### Statistics and reproducibility

Experiments were repeated independently in two monkeys and data were combined for subsequent analysis after we confirmed that behaviour was similar across monkeys (Extended Data Fig. 1). Tests were not corrected for multiple comparisons unless otherwise specified. Nonparametric tests were performed using 1,000 iterations; therefore, exact  $P$  values are specified when  $P \geq 0.001$ .

Analyses were performed in MATLAB (Mathworks).

### Mixture modelling of behavioural reports

Behavioural errors on delayed estimation tasks are thought to be due to at least three sources of errors<sup>12,13</sup>: imprecise reports of the cued stimulus, imprecise reports of the uncued stimulus, and random guessing (that is, from ‘forgotten’ stimuli). To estimate the contribution of each of these sources of error, we used a three-component mixture model to model behavioural reports<sup>13</sup>:

$$p(\hat{\theta}) = (1 - \gamma - B)\phi_o(\hat{\theta} - \theta) + \gamma\frac{1}{2\pi} + B\frac{1}{m}\phi_o(\hat{\theta} - \theta^*)$$

in which  $\theta$  is the colour value of the cued stimulus in radians,  $\hat{\theta}$  is the reported colour value,  $\theta^*$  is the colour value of the uncued stimulus,  $\gamma$  is the proportion of trials on which subjects responded randomly (that is, probability of guessing,  $P(\text{guess})$ ),  $B$  is the proportion of trials on which subjects reported the colour of the uncued stimulus (that is, probability of ‘swapping’,  $P(\text{swap})$ ), and  $\phi_o$  is a von Mises distribution with a mean of zero and a standard deviation  $\sigma$  (inverse precision). All parameters were estimated using the Analogue Report Toolbox (<https://www.paulbays.com/toolbox/index.php>). Bootstrapped distributions of the maximum likelihood values of the free parameters  $\gamma$ ,  $B$ , and  $\sigma$

were generated by fitting the mixture model independently to the behavioural data from each session ( $n = 23$ ) and then resampling the best-fitting parameter values with replacement across sessions. In this way, the distribution shows the uncertainty of the mean parameters across sessions.

As noted in the main text, if the monkey was able to select an item from memory earlier in the trial, then this reduced the error in the monkey’s behavioural response (Extended Data Fig. 1e). Behavioural modelling showed that earlier cues improved the precision of memory reports (Extended Data Fig. 1e,  $\beta = 3.95 \pm 1.88$  s.e.m.,  $P = 0.012$ , bootstrap) but did not significantly change the probability of forgetting (that is, random responses;  $\beta = 0.03 \pm 0.03$  s.e.m.,  $P = 0.126$ , bootstrap). Furthermore, we found that memory reports were more accurate in the pro condition than in the retro condition (Fig. 1b, c). Here, behavioural modelling showed that the improvement on pro trials was due to an increase in the precision of memory reports and a reduction in forgetting (that is, fewer random reports) (Extended Data Fig. 1d).

### Entropy of report distributions

To quantify whether colour reports were more clustered than expected by chance, we used a simple clustering metric<sup>30</sup>. This metric relies on the fact that entropy is maximized for uniform probability distributions. By contrast, probability distributions with prominent peaks will have lower entropy. Because the target colours are drawn from a circular uniform distribution, the entropy of the targets  $H(\theta)$  will be relatively high. If responses are clustered, their entropy  $H(\hat{\theta})$  will be relatively low. Taking the difference of these two values yields a clustering metric  $C$ . Negative values of  $C$  indicate greater clustering:  $C = H(\hat{\theta}) - H(\theta)$ , in which  $H(x) = -\sum_{x=1}^{360} f(x) \log_2 f(x) dx$ . The significance of the clustering metric versus zero was assessed with a bootstrapping process that randomly resampled trials with replacement.

### Calculation of cued location $d'$

We used  $d'$  to describe how each neuron’s firing rate was modulated by cuing condition (‘upper’ or ‘lower’), defined as:

$$d' = \frac{\mu_{\text{upper}} - \mu_{\text{lower}}}{\sqrt{\frac{1}{2}(\sigma_{\text{upper}}^2 + \sigma_{\text{lower}}^2)}}$$

in which  $\mu_{\text{upper}}$  and  $\mu_{\text{lower}}$  are a neuron’s mean firing rate on trials in which the upper or lower stimulus was cued as task relevant, respectively, and  $\sigma_{\text{upper}}^2$  and  $\sigma_{\text{lower}}^2$  are the variance in firing rate across trials in each condition.  $d'$  was either computed using trials pooled across all retro trials (Extended Data Fig. 2b) or calculated separately for each of the three block types (Extended Data Fig. 4a; pro with cue set 1, retro with cue set 1, and retro with cue set 2, see above). This analysis included all neurons that were recorded for at least ten trials per cued location. The significance of each neuron’s  $d'$  (Extended Data Fig. 2b) was assessed by comparing to a null distribution of values generated by randomly permuting location labels (upper or lower) across trials (1,000 iterations). To test whether a region had more significant neurons than expected by chance, the percentage of significant neurons was compared to that expected by chance (the  $\alpha$ -level, 5%).

To understand whether cells displayed similar selectivity across cue sets and task conditions, we computed a ‘selection’ correlation, measured as the Pearson’s correlation coefficient between the  $d'$  to retro cue set 1 and the  $d'$  to retro cue set 2, and a ‘generalization’ correlation, measured as the Pearson’s correlation coefficient between pro cue set 1 and retro cue set 2 (Extended Data Fig. 4a–c). Significance against zero was tested by randomly resampling cells with replacement.

### Classification of cued location

We used linear classifiers to quantify the amount of information about the location of the cued stimulus (upper or lower) in the population

of neurons recorded from each brain region (Fig. 2b, c). This analysis included all neurons that were recorded during at least 60 trials for each cueing condition (upper or lower) in each block type (pro with cue set 1, retro with cue set 1, and retro with cue set 2, see above). On each of 1,000 iterations, 60 trials from each cueing condition and block type were sampled from each neuron with replacement. The firing rate from those trials, locked to cue onset, was assembled into a pseudo-population by combining neurons across sessions such that pseudo-trials matched both block and cue condition. For each time step, a logistic regression classifier (as implemented by `fitlinear.m` in MATLAB) with L2 regularization ( $\lambda = 1/60$ ) was trained to predict the cueing condition (upper or lower) using pseudo-population data from one block (for example, retro with cue set 1) and tested on held-out data from another block (for example, retro with cue set 2). Classification accuracy (proportion of correctly classified trials) was averaged across reciprocal tests (for example, train on retro with cue set 2, test on retro with cue set 1).

We used a randomization test to test for significant differences in the onset time of above-chance classification accuracy between regions. For each pair of regions, we computed the difference in time of first significance ('lag',  $P < 0.05$ , using the bootstrap procedure described above). To generate a null distribution of lags, we randomly permuted individual neurons between the two regions (without changing the size of the population associated with each region) and then repeated the above bootstrap procedure to determine the lag in above-chance classification for each permuted dataset. One thousand random permutations were used for each pair of regions. Significance was assessed by computing the proportion of lags in the null distribution that were greater than the observed lag. This randomization procedure controls for differences in the number of features (neurons) across regions, so differences in the number of neurons recorded across regions cannot explain our results.

To assess the discriminability of the upper and lower pro conditions (Extended Data Fig. 4e), we calculated the tenfold cross-validated classification accuracy (averaged across folds). To provide an estimate of variability we repeated this analysis 1,000 times, each time with a different partition of trials into training and testing sets.

## Neuron dropping analyses for classification of cued location

To further test whether classification performance depended on the number of neurons recorded in each region, we performed 'neuron dropping' analyses<sup>33</sup> (Extended Data Fig. 2). To do this, we repeated the classification procedure described above, but limited the analysis to subsets of neurons drawn from the full population of neurons recorded in each region ( $n = 1,000$  iterations per subset size). In the first version of this analysis, the neurons that composed each subpopulation were drawn at random (Extended Data Fig. 2c). In the second version of this analysis, the neurons that composed each subpopulation were drawn at random, subject to the condition that they displayed a significant evoked response to the presentation of the cue (Extended Data Fig. 2d). Specifically, across trials, neurons with evoked responses were taken as those with a higher mean firing rate during the 500-ms epoch after the cue compared to the 300-ms epoch before the cue (one-tailed  $t$ -test). In the third version of this analysis, neurons were added to the analysis in a fixed order determined by their ability to support classification (Extended Data Fig. 2e, f). For the selection classifier (which was trained to discriminate the cued location on retro cue set 1 trials and tested on retro cue set 2 trials, and vice-versa), neurons entered the analysis based on the magnitude of their  $d'$  values for both retro cue sets. To quantify this, we projected the  $d'$  values for the two cue sets onto the identity line (schematized in Extended Data Fig. 2e) and took the absolute value of the resulting vector. Cells with large absolute projection values entered the analysis first. Our ordering procedure for the generalization classifier was the same as for the selection classifier, except that it was based on pro cue set 1 and retro cue set 2  $d'$ , as these were the training and testing sets for this classifier.

For each subpopulation of neurons in each of these analyses, we measured four statistics: (1) selection classification accuracy after cue onset (300 ms post-cue); (2) generalized classification accuracy after cue onset (300 ms post-cue); (3) time to 55% selection classification accuracy; and (4) time to 55% generalized classification accuracy.

When subpopulations were drawn at random from all neurons in each region or all neurons that displayed an evoked response (Extended Data Fig. 2c, d), dropping curves for each of these statistics were well described by two-parameter power functions. Power functions were fit using the Matlab function `fit.m` and 95% prediction intervals for each statistic at the maximum population size recorded in LPFC were generated using `predint.m`. The distance of the measured statistic in LPFC from these predicted values (in units of standard error of the prediction interval) were measured and used to calculate  $P$  values.

When subpopulations were drawn in a fixed order (Extended Data Fig. 2e, f), dropping curves for each of these statistics were well described by linear functions. Linear functions were fit using the Matlab function `fit.m` and 95% confidence intervals for linear fits at each measured value were generated using `predint.m`. Subpopulations that never reached 55% classification accuracy were excluded from curve fitting for statistics 3 and 4.

Finally, to assess the discriminability of visual information in each region we trained two classifiers to discriminate either the two upper cues or two lower cues (on retro trials). Classification accuracy was averaged across tenfold cross-validated sets (Extended Data Fig. 2g, h). The accuracies of both the upper-cue and lower-cue classifiers were then averaged to estimate the amount of information about low-level visual features of the cue while holding other factors constant (for example, cued location). We then computed neuron dropping curves for (1) accuracy early after cue onset (300 ms post-cue) and (2) time to 55% classification accuracy, as above (Extended Data Fig. 2h).

## Signal and noise for classification of cued location

To assess whether classification performance was driven by increases in signal, decreases in noise, or both, we analysed the distribution of classifier confidence for 'upper' and 'lower' test trials (Extended Data Fig. 3a). Classifier confidence was quantified as the probability that a given test trial was an upper trial, as estimated by the trained model. Signal was quantified as the distance between the means of the confidence distributions for upper and lower trials and noise was estimated as the average standard deviation of the confidence distributions. Repeating these calculations for each of the 1,000 resamples yielded bootstrapped distributions of values (Extended Data Fig. 3b).

## Noise correlations and variance-to-mean ratio during cue epoch

To determine whether differences in variance and covariance might drive differences in classification performance across regions, we calculated variance-to-mean ratios and noise correlations for single-neuron firing rates around the time of the selection cue.

Variance-to-mean ratio across trials was calculated by first calculating each neuron's trial-wise firing rate during the period after cue onset (0–500 ms post-cue). Next, for each trial type (pro cue set 1 upper, pro cue set 1 lower, retro cue set 1 upper, retro cue set 1 lower, pro cue set 2 upper, pro cue set 2 lower), we divided the variance of these firing rates across trials by their mean. Finally, we took the average of these variance-to-mean ratios across the six trial types (Extended Data Fig. 3d).

Calculation of noise correlations also began by first calculating each neuron's trial-wise firing rate during the period after cue onset (0–500 ms post-cue). Next, for each trial type (pro cue set 1 upper, pro cue set 1 lower, retro cue set 1 upper, retro cue set 1 lower, pro cue set 2 upper, pro cue set 2 lower), we adjusted each neuron's pool of firing rates to have a mean of zero. Finally, we computed the average correlation between the mean-zeroed firing rates of all pairs of neurons within a region, and then averaged these average correlation values across the six trial



types (Extended Data Fig. 3c). As expected, given our pseudopopulation approach, noise correlation values were low and did not differ across regions.

### Quantification of colour information

We adapted previous work<sup>34</sup> to define a colour modulation index ( $MI_{\text{colour}}$ ) that describes how each neuron's firing rate was modulated by the colours of the remembered stimuli. Critically, this statistic avoids strong assumptions about the structure of tuning curves (for example, it does not assume unimodal tuning). After dividing colour space into  $N = 8$  bins,  $MI_{\text{colour}}$  is defined as:

$$MI_{\text{colour}} = \frac{\sum_{c=1}^N z_c \log(Nz_c)}{\log(N)}$$

in which  $z_c$  is a neuron's normalized mean firing rate  $r_c$  across trials evoked by colours in the  $c$ -th bin:

$$z_c = \frac{r_c}{\sum_{c=1}^N r_c}$$

$MI_{\text{colour}}$  is a normalized entropy statistic that is 0 if a neuron's mean firing rate is identical across all colour bins and 1 if a neuron fires only in response to colours from one bin. To control for differences in average firing rate and number of trials across neurons, we z-scored this metric by subtracting the mean and dividing by the standard deviation of a null distribution of MI values. To generate this null distribution, the colour bin labels were randomly shuffled across trials and the MI statistic was recomputed (1,000 times per neuron).

Z-scored colour modulation indices were computed separately for each time point, trial type (pro or retro), and stimulus type (selected, non-selected, attended or non-attended) (Fig. 3b, Extended Data Fig. 6). This analysis included neurons that were recorded for at least ten trials in each of these conditions. Selectivity for colour was computed without respect to the spatial location of the stimulus (upper or lower). Computing selectivity for colours presented only at a neuron's preferred location did not qualitatively change the results. Z-scored modulation indices were compared to zero or across conditions by  $t$ -test (Fig. 3b). We corrected for multiple comparisons over time using a cluster-correction<sup>35</sup>. In brief, the significance of contiguous clusters of significant  $t$ -tests was computed by comparing their cluster mass (the sum of the  $t$ -values) with what would be expected by chance (randomization test). In addition, to summarize changes in selected and non-selected colour information after cue onset, we averaged colour information for each neuron in two time periods ( $-300$  to  $0$  ms pre-cue and  $200$  to  $500$  ms post-cue) and tested the difference between these values (post-pre) against zero by bootstrapping the mean difference in colour information across neurons (Extended Data Fig. 7a).

To determine whether a neuron displayed significant selectivity for the colour at one particular location (upper/lower), we calculated the z-scored information about the cued colour at each time point over the interval from  $0$  to  $2.5$  s after stimulus onset independently for each location. Colour selectivity was measured across all conditions, including pro, retro, and single-item trials. As described above, we used a cluster correction to correct for multiple comparisons across time. Neurons with significant colour selectivity ( $P < 0.05$ ) at any point during this interval were deemed colour selective. Binomial tests compared the proportion of neurons with significant colour selectivity for at least one of the two locations to a conservative null proportion of 10% (for two tests with an  $\alpha$  of 0.05, one test for each location).

To determine whether independent populations of LPFC neurons encoded the upper and lower colours during the pre-cue period of retro trials, we counted the number of neurons with significant cluster-corrected selectivity during the 500-ms period before cue onset. Of the 607 LPFC neurons that entered the analysis, 112 (18.5%)

carried information about the upper colour and 99 (16.3%) carried information about the lower colour. Of these, 35 (5.8%) carried information about both the upper and lower colour. A binomial test compared this proportion (5.8%) to that expected by random assignment of top- and bottom-selectivity (that is,  $18.5\% \times 16.3\% \approx 3.0\%$ ). To visualize selectivity in a non-binary manner, we also plotted the distribution of z-scored information about the colour of each item for all LPFC neurons, averaged during the 500-ms pre-cue period (Extended Data Fig. 8a).

In addition to the z-scored colour modulation index, we also quantified colour selectivity using per cent explained variance (PEV) (Extended Data Fig. 5b). As with the z-scored colour modulation index, firing rates for each time point were binned by the colour of the stimulus of interest (selected or non-selected) into eight colour bins. A linear model with a constant term and eight categorical predictors (one for each colour bin) was then constructed to predict firing rates (using `fitlm.m` in MATLAB). PEV was then calculated as the  $r^2$  of the fit model  $\times 100$ . To avoid inflated PEV values due to overfitting, we subtracted the mean PEV during the 200-ms epoch before stimulus onset. The resulting traces were analysed using cluster-corrected  $t$ -tests, as described above. The results were similar to those obtained with the colour modulation index.

### Quantification of reported colour information

To quantify the amount of information each neuron carried about the monkey's reported colour, we followed the same approach as for stimulus colour, except that responses were binned by the colour reported by the monkey rather than by the colour of the cued or uncued stimulus (Fig. 3).

### Modulation of colour information by task and behavioural performance

To compare the amount of colour information in firing rates across the pro and retro conditions (Extended Data Fig. 6b, c), we computed the z-scored colour modulation indices as described above for each of the four conditions of interest (selected, non-selected, attended, and non-attended colours). Trial counts were matched across these four conditions to avoid biases in the colour information statistic. To assess relative information about cued (selected and attended) and uncued (non-selected and non-attended) colour information, we computed the difference in colour information between each pair of conditions, for each neuron. The average difference across all neurons was then tested against zero, using the cluster correction described above to correct for multiple comparisons across time<sup>35</sup>.

To compare the amount of colour information in firing rates when behavioural performance was relatively accurate or inaccurate (Extended Data Fig. 7), we divided retro trials into two groups according to the accuracy of the behavioural report. Trials within each session were split by the median accuracy for that session. Z-scored colour modulation indices were computed separately for each split-half of trials. As above, the same number of trials were used for all four conditions (more or less accurate  $\times$  selected or non-selected). In addition, to quantify the effect of selection, the difference in colour information for selected and non-selected colours was computed for each group of trials separately (more or less accurate). This selected-non-selected difference was then tested against zero to measure the effect of selection and tested between the two groups of trials to measure the effect of behavioural accuracy. Comparisons were done with a  $t$ -test across all neurons, with cluster correction to correct for multiple comparisons across time<sup>35</sup>.

### Measuring the angle between upper and lower colour planes

As described in the main text, we were interested in understanding the geometry of mnemonic representations of colour across the two possible stimulus locations (upper or lower). To explore this, we examined the response of the population of neurons as a function of the

# Article

colour and location of the stimulus of interest (either cued or uncued). The fidelity of these population representations depended on the behavioural performance of the monkey. Therefore, for all principal component analyses, we divided trials on the basis of the accuracy of the behavioural report (median split for each session, as above) and separately analysed trials with lower angular error (higher accuracy, Fig. 4) and higher angular error (lower accuracy, Extended Data Fig. 9f).

Trials were sorted into  $B=4$  colour bins and  $L=2$  locations (top or bottom), yielding  $B \times L = M$ , eight total conditions. To visualize these population representations, we projected the population vector of mean firing rates for each of these eight conditions into a low-dimensional coding subspace (Fig. 4a, Extended Data Fig. 9a, b, similar to previous work<sup>36</sup>). For each time step, we defined a population activity matrix  $X$  as an  $M \times N$  matrix, in which  $M$  is the number of conditions (eight) and  $N$  is the number of neurons:

$$X = \begin{bmatrix} \mathbf{r}(c_{1,L}) - \bar{\mathbf{r}} \\ \vdots \\ \mathbf{r}(c_{B,L}) - \bar{\mathbf{r}} \end{bmatrix}$$

Here,  $\mathbf{r}(c_{B,L})$  is the mean population vector (across trials) for the condition corresponding to colour bin  $B$  and location  $L$ , and  $\bar{\mathbf{r}}$  is the mean population vector across the  $M$  conditions (that is, the mean of each column is zero).

The principal components of this matrix were identified by decomposing the covariance matrix  $C$  of  $X$  using singular value decomposition (as implemented by `pca.m` in MATLAB):  $C = PDP^T$ , in which each column of  $P$  is an eigenvector of  $C$  and  $D$  is a diagonal matrix of corresponding eigenvalues. We constructed a reduced ( $K=3$ )-dimensional space whose axes correspond to the first  $K$  eigenvectors of  $C$  (that is, columns of  $P$ ,  $P_K$ , assuming eigenvectors are ordered by decreasing explained variance). These first three eigenvectors explained an average of 65% of the variance in the mean population response across all examined time points. We then projected the population vector for a given condition into this reduced dimensionality space:  $z_K = P_K^T(\mathbf{r}(c_{B,L}) - \bar{\mathbf{r}})$ , in which  $z_K$  is the new coordinate along axis  $K$  in the reduced dimensionality space.

We observed that, when visualized in the reduced-dimensionality space, the population representations for each colour bin  $B$  within a given location  $L$  tended to lie on a plane, referred to as the ‘colour plane’ in the main text (Fig. 4a). To identify the best-fitting plane, we defined a new population activity matrix  $Y_L$  for each location  $L$  with dimensions  $B \times K$ :

$$Y_L = \begin{bmatrix} \mathbf{z}(c_{1,L}) - \bar{\mathbf{z}}_L \\ \vdots \\ \mathbf{z}(c_{B,L}) - \bar{\mathbf{z}}_L \end{bmatrix}$$

in which  $\mathbf{z}(c_{B,L})$  is the population vector for the condition corresponding to colour bin  $B$  and location  $L$  in the reduced dimensionality space, and  $\bar{\mathbf{z}}_L$  is the mean population vector across colour bins for that location (that is, the mean of each column is zero). The principal components of this matrix were calculated in the same manner as above and the first two principal components were the vectors that defined the plane-of-best-fit to the points defined by the rows of  $Y_L$ . These planes explained more than 97% of the variance of each set of points in the 3D subspace.

If the vectors defining the plane-of-best-fit for the upper item are  $\mathbf{v}_1$  and  $\mathbf{v}_2$  and those for the lower item are  $\mathbf{v}_3$  and  $\mathbf{v}_4$ , then the cosine of the angle between these two colour planes can be calculated as:

$$\cos(\theta) = (\mathbf{v}_1 \times \mathbf{v}_2) \times (\mathbf{v}_3 \times \mathbf{v}_4)$$

For all analyses, population vectors were based on pseudo-populations of neurons combined across sessions. Pseudo-populations were created by matching trials across sessions according to the colour and location

of the stimulus of interest (either cued or uncued), as described above (and following previous work<sup>27</sup>). This analysis included only neurons that were recorded for at least ten trials for each conjunction of colour and location. Confidence intervals of  $\cos(\theta)$  were calculated using a bootstrapping procedure. On each of 1,000 iterations, 10 trials from each of the 8 conditions were sampled from each neuron with replacement. The average firing rates across these sampled trials provided the mean population vector for that condition on that iteration. To assess how  $\cos(\theta)$  changed around cue onset (Fig. 4b, Extended Data Fig. 9f), we used a logistic regression model of the form:

$$\cos(\theta) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 t))}$$

in which  $t$  is time relative to cue onset. This model was fit to values of  $\cos(\theta)$  computed at each time point in the interval from 500 ms before to 1,000 ms after cue onset on each bootstrap iteration (described above). This yielded a bootstrapped distribution of  $\beta_1$  estimates that could be compared to zero or across the two groups of trials with more and less accurate behavioural responses (Extended Data Fig. 9f).

## Defining the colour subspaces for the upper and lower items in the full-dimensional space

To define the colour subspace in the full neuron-dimensional space, we defined  $B=4 \times N$  mean population activity matrices  $L$  in the full space:

$$W_L = \begin{bmatrix} \mathbf{r}(c_{1,L}) - \bar{\mathbf{r}}_L \\ \vdots \\ \mathbf{r}(c_{B,L}) - \bar{\mathbf{r}}_L \end{bmatrix}$$

The colour subspace was defined as the first two principal components of  $W_L$ .

These subspaces were used for two analyses. First, we projected the population vectors of colour responses from one item into the colour subspace for the other item (Fig. 4d). For example, the population vector response to colours of the upper item were projected into the colour subspace of the lower item, defined as the first two principal components of  $W_{\text{lower}}$ , and vice versa (Fig. 4d). Second, by defining the colour subspace of each item at different time points  $t_i$ , we could examine how colour representations evolved during the trial (Fig. 4e, Extended Data Fig. 8c, d).

## Measuring the separability of colours in a subspace

Next, we were interested in quantifying the separability of colours in a given subspace. As seen in Fig. 4d, e, the population representation of the four colour conditions, projected into the subspace, form the vertices of a quadrilateral with the edges of the quadrilateral connecting adjacent colours on the colour wheel (for example, Fig. 4d). To measure separability of the colours, we computed the area of this quadrilateral (polyarea.m function in MATLAB). Bootstrapped distributions of these area estimates were obtained by resampling trials with replacement from each condition before re-computing  $W_L$ .

## Similarity of transforms for the upper and lower stimulus

We were interested in testing whether the transformation of selected (‘cued’) items was the same as non-selected (‘uncued’) items (on retro trials). To this end, we examined how the population representation for the colour of the selected and non-selected stimuli changed over time. For both a pre-cue (150 to 350 ms post-stimulus offset) and post-cue (–200 to 0 pre-target onset) time epoch, we defined an  $N \times B$  population activity matrix  $A$ , in which  $N$  is the number of neurons,  $B=4$  colour bins, and the elements of the matrix reflect the mean firing rate of each neuron across trials in which the colour of the stimulus of interest fell in colour bin  $b$ .

We computed  $A_{\text{pre}}$  and  $A_{\text{post}}$  separately for four different stimulus types of interest: cued upper stimuli, cued lower stimuli, uncued upper stimuli, and uncued lower stimuli. Then, for each stimulus type, we identified the  $N \times N$  matrix  $X$  that transformed the pre-cue representation to its post-state:

$$A_{\text{post,cued\_upper}} = X_{\text{cued\_upper}} A_{\text{pre,cued\_upper}}$$

$$A_{\text{post,cued\_lower}} = X_{\text{cued\_lower}} A_{\text{pre,cued\_lower}}$$

$$A_{\text{post,uncued\_upper}} = X_{\text{uncued\_upper}} A_{\text{pre,uncued\_upper}}$$

$$A_{\text{post,uncued\_lower}} = X_{\text{uncued\_lower}} A_{\text{pre,uncued\_lower}}$$

To assess how similar these transforms were, we applied transforms from one condition (for example, cued upper) to held-out (split half) pre-cue neural data ( $A_{\text{pre}}^{\text{withheld}}$ ) from a different condition (for example, cued lower) and compared how similar the predicted post-cue data ( $A_{\text{post}}^{\text{predicted}}$ ) were to the actual (held-out) post-cue data ( $A_{\text{post}}^{\text{withheld}}$ ). Reconstruction error was measured as the Euclidean distance between the predicted and actual population vectors ( $A_{\text{post}}^{\text{withheld}} - A_{\text{post}}^{\text{predicted}}$ ), averaged across all colours. Low reconstruction error indicates similar transforms.

This procedure allowed us to determine how similar the transforms were across locations and cue types by testing whether the transformation, defined in one condition for one item, generalized to another condition and/or another item. For example, for the ‘cued upper’ condition, the reconstruction errors of different forms of generalization were computed as follows:

$$\begin{aligned} &\text{Error(same condition, same item)} \\ &= f(A_{\text{post,cued\_upper}}^{\text{withheld}} - X_{\text{cued\_upper}} A_{\text{pre,cued\_upper}}) \end{aligned}$$

$$\begin{aligned} &\text{Error(same condition, different item)} \\ &= f(A_{\text{post,cued\_upper}}^{\text{withheld}} - X_{\text{uncued\_lower}} A_{\text{pre,cued\_upper}}) \end{aligned}$$

$$\begin{aligned} &\text{Error(different condition, same item)} \\ &= f(A_{\text{post,cued\_upper}}^{\text{withheld}} - X_{\text{uncued\_upper}} A_{\text{pre,cued\_upper}}) \end{aligned}$$

$$\begin{aligned} &\text{Error(different condition, different item)} \\ &= f(A_{\text{post,cued\_upper}}^{\text{withheld}} - X_{\text{cued\_lower}} A_{\text{pre,cued\_upper}}) \end{aligned}$$

in which  $f$  is the mean root sum of squares across columns (that is, the mean Euclidean distance between the actual and reconstructed population vectors for each colour bin  $b$ ). Similar reconstruction errors were estimated for the other three conditions (cued lower, uncued upper, and uncued lower).

Applying the estimated transform to held-out data from the same condition and the same item provides a lower bound on reconstruction error due to variance across trials and indicates whether the transformations are stable within a condition. Applying transforms to the response to the other item in the same cuing condition (same condition, different location) allows us to test whether the selected and non-selected items are transformed in similar ways by comparing reconstruction error to (1) chance and (2) the error within condition and within item (same condition, same item). Finally, to control for any similarity in transforms due to a non-condition-specific effect of the cue (for example, time during the task), we can apply transforms based on items in the other cuing condition, either to the same item (different condition, same item) or the other item (different condition, different item).

We computed the four types of reconstruction error by averaging across all four conditions of interest (cued upper, cued lower, uncued upper, uncued lower). To estimate the distribution of reconstruction error, we bootstrapped with replacement across trials. Chance levels of reconstruction error were estimated by repeating the bootstrapping procedure but permuting the condition label (cued upper, cued lower, uncued upper, uncued lower) assigned to each colour population vector.

### Correlation of colour representations

We wanted to understand how similarly colour was represented across the upper and lower locations over the course of the trial. To investigate this, we binned retro or pro trials according to the colour and location of the stimulus of interest (cued or uncued) and then randomly partitioned into two halves. These split halves were used to estimate the degree of noise in the data (Extended Data Fig. 10b–d, described below). Specifically, trials were sorted into  $B=4$  colour bins,  $L=2$  locations (top or bottom), and  $H=2$  halves, yielding  $B \times L \times H = M$  total conditions. For each of these conditions, at a given time point of interest, we computed the average population vector  $\mathbf{r}(c_{B,L,H})$ .

We then computed the average correlation between each population vector and the population vectors corresponding to the same colour bin at the other location (Fig. 4c, Extended Data Figs. 9c, 10b–d):

$$\rho_{\text{cross}} = \frac{1}{B2H} \sum_{i=1}^H \sum_{j=1}^H \sum_{b=1}^B \text{corr}(\mathbf{r}(c_{b,1,i}) - \langle \mathbf{r}(c_{B,1,i}) \rangle_B, \mathbf{r}(c_{b,2,j}) - \langle \mathbf{r}(c_{B,2,j}) \rangle_B)$$

in which  $\langle \cdot \rangle_B$  is the average across the set of colour bins  $B$ . In other words, for each set of  $B$  population vectors corresponding to a particular half of the data  $H$  and location  $L$ , we subtracted the mean across bins to centre the vector endpoints around zero. Thus,  $\rho_{\text{cross}}$  quantifies to what extent colour representations are similarly organized around their mean across the two locations.

To obtain an upper bound on potential values of  $\rho_{\text{cross}}$  given the degree of noise in the data, we also computed the average correlation of each population vector with itself across the two halves:

$$\rho_{\text{self}} = \frac{1}{BL} \sum_{b=1}^B \sum_{l=1}^L \text{corr}(\mathbf{r}(c_{b,1,l}) - \langle \mathbf{r}(c_{B,1,l}) \rangle_B, \mathbf{r}(c_{b,1,2}) - \langle \mathbf{r}(c_{B,1,2}) \rangle_B)$$

Finally, to understand how similarly colour was represented across the two cueing conditions, trials were sorted into  $B=4$  colour bins,  $L=2$  locations (top or bottom), and  $C=2$  cuing conditions (pro and retro). For each of these conditions, at a given time point of interest, we computed the average population vector  $\mathbf{r}(c_{B,L,C})$ . We then computed the average correlation between each population vector and the population vectors corresponding to the same colour bin at either the same or different location in the other task (Extended Data Fig. 10f):

$$\rho_{\text{att,sel}} = \frac{1}{B2L} \sum_{i=1}^L \sum_{j=1}^L \sum_{b=1}^B \text{corr}(\mathbf{r}(c_{b,i,1}) - \langle \mathbf{r}(c_{B,i,1}) \rangle_B, \mathbf{r}(c_{b,j,2}) - \langle \mathbf{r}(c_{B,j,2}) \rangle_B)$$

To compare the similarity of colour representations on retro trials to pre-target pro colour representations, we computed this correlation between (1) the response on pro trials, for all time points falling within the interval from  $-300$  ms to  $0$  ms before the onset of the response wheel, and (2) the response on retro trials at two different time points: before selection (from  $-300$  to  $0$  ms before the cue) and after selection (from  $-300$  to  $0$  ms before the onset of the response wheel). Correlation was measured between each time point across windows and then averaged across all pairs of time points.

As above, population vectors were pseudo-populations of neurons combined across sessions, in which trials across sessions were matched according to colour bin and location<sup>27</sup>. This analysis included only neurons that were recorded for at least ten trials for each conjunction of

# Article

colour and location. Confidence intervals for  $\rho_{\text{cross}}$ ,  $\rho_{\text{self}}$ , and  $\rho_{\text{att,sel}}$  were calculated with a bootstrap. On each of 1,000 iterations, and for each neuron and condition (colour–location–half conjunction), the entire population of trials in that condition was resampled with replacement. The average firing rates across these sampled trials provided the mean population vector for that condition on that iteration. As with principal components analyses, we divided trials on the basis of the accuracy of the behavioural report (median split of trials for each session) and the presented results reflect analysis of trials with lower angular error, unless otherwise noted.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Data supporting all figures are included with the manuscript. Raw electrophysiological and behavioural data are available from the corresponding author upon reasonable request. Source data are provided with this paper.

## Code availability

Behavioural code and custom Matlab analysis functions are publicly available at <https://github.com/buschman-lab/>. All other code is available from the authors upon reasonable request.

31. Bruce, C. J. & Goldberg, M. E. Primate frontal eye fields. I. Single neurons discharging before saccades. *J. Neurophysiol.* **53**, 603–635 (1985).
32. Rolston, J. D., Gross, R. E. & Potter, S. M. Common median referencing for improved action potential detection with multielectrode arrays. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2009**, 1604–1607 (2009).
33. Wessberg, J. et al. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408**, 361–365 (2000).
34. Tort, A. B. L., Komorowski, R., Eichenbaum, H. & Kopell, N. Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *J. Neurophysiol.* **104**, 1195–1210 (2010).
35. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
36. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).

**Acknowledgements** We thank B. Morea and H. Weinberg-Wolf for assistance with monkeys; S. Tafazoli for assistance with microstimulation; F. Bouchacourt, C. Jahn, A. Libby, C. MacDowell, S. Tafazoli, M. Uchimura, and S. Henrickson for feedback; and the Princeton Laboratory Animal Resources staff for support. This work was supported by NIMH R01MH115042 (T.J.B.) and an NDSEG Fellowship (M.F.P.).

**Author contributions** T.J.B. conceived the project. M.F.P. trained the monkeys, collected the data, and analysed the data with supervision from T.J.B. T.J.B. and M.F.P. wrote the paper.

**Competing interests** The authors declare no competing interests.

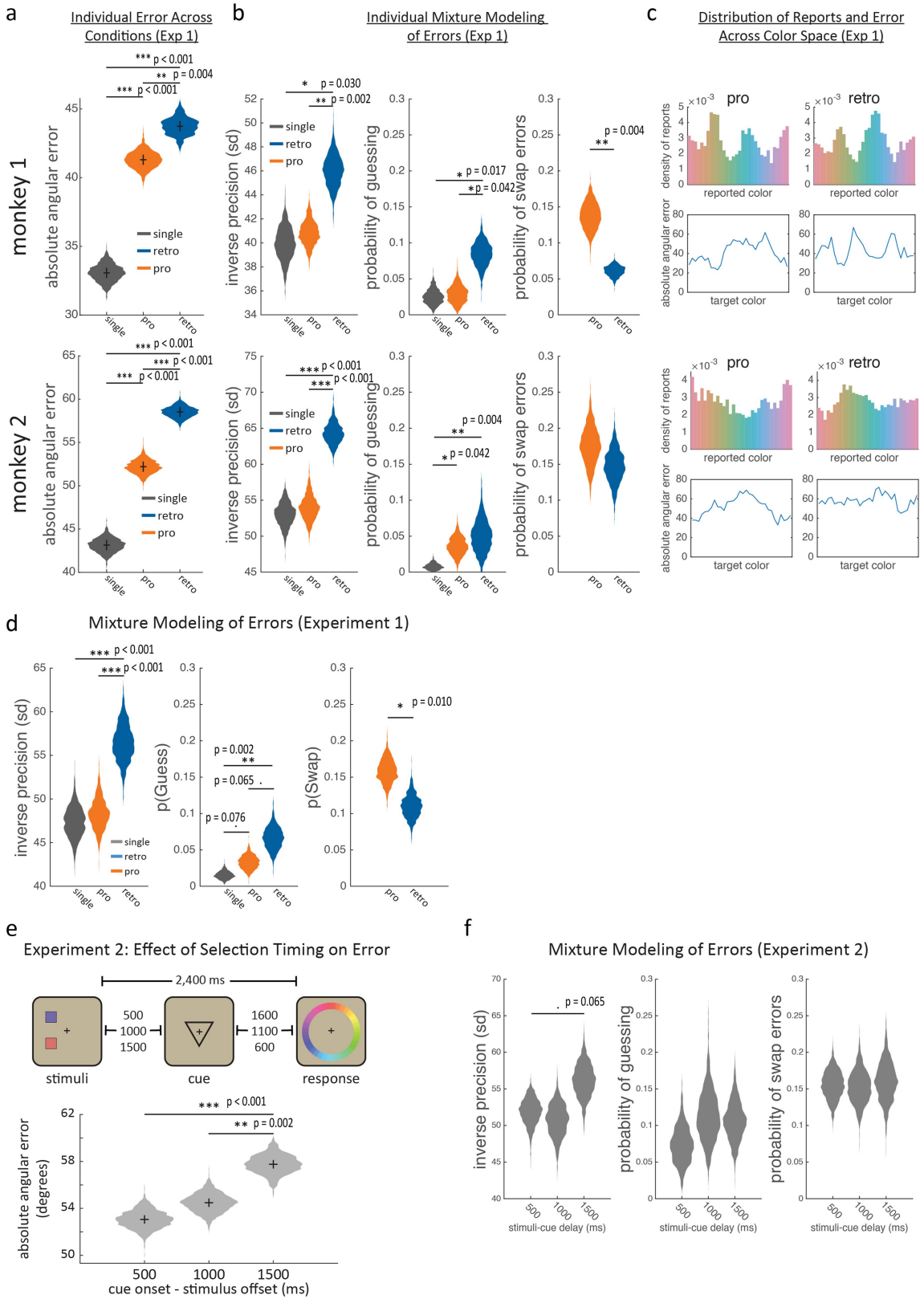
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-021-03390-w>.

**Correspondence and requests for materials** should be addressed to T.J.B.

**Peer review information** Nature thanks Tirin Moore and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

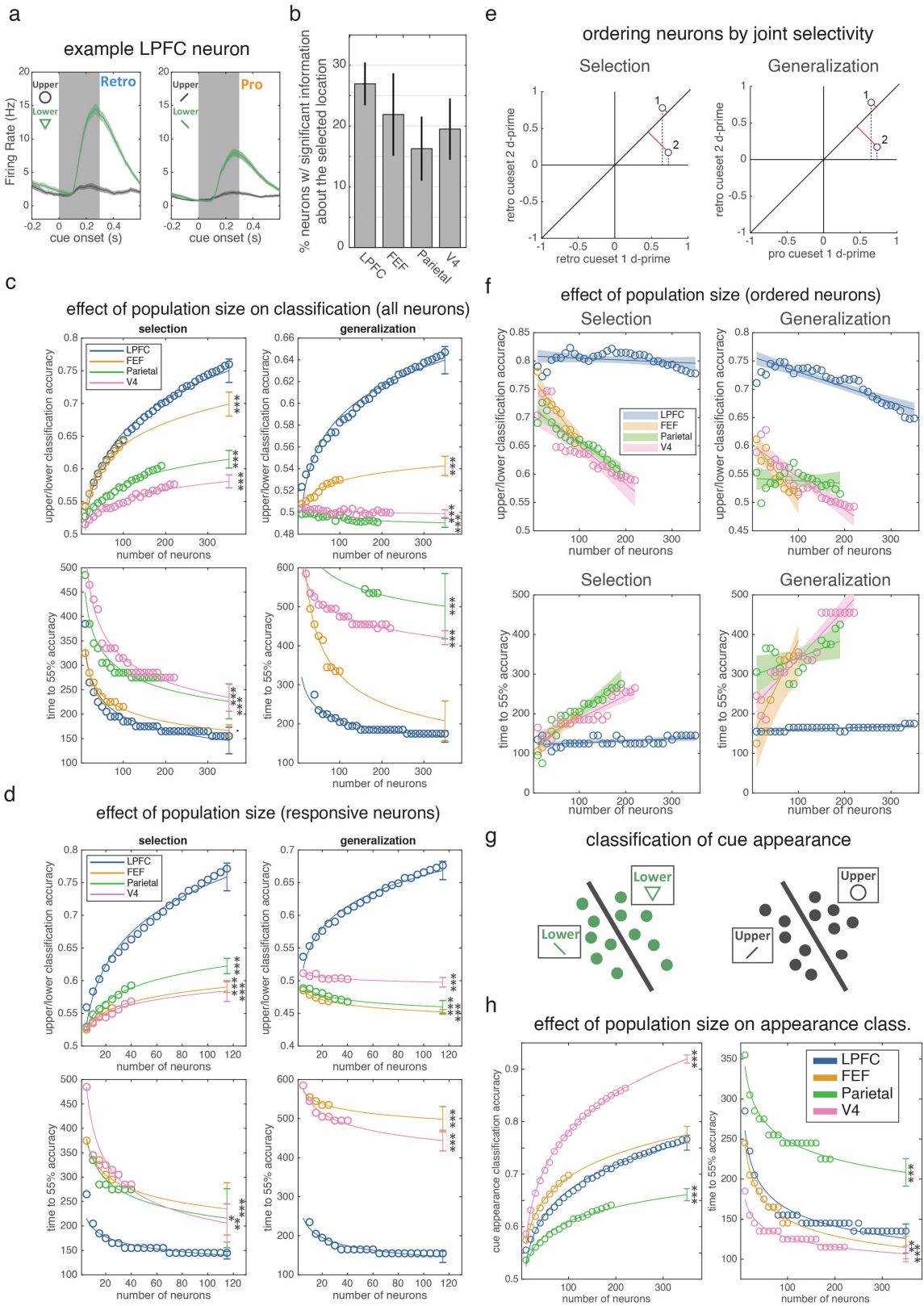


**Extended Data Fig. 1** | See next page for caption.

# Article

**Extended Data Fig. 1 | Behaviour was consistent across monkeys and selection mitigated the decay of memories over time.** **a, b**, Mean absolute angular error (**a**) and mean mixture model parameter fits (**b**) in the main experiment (experiment 1) (Fig. 1a) for each monkey (Methods). Violin plots depict bootstrapped distribution across sessions ( $n = 10$  for monkey 1 and  $n = 13$  for monkey 2). Lines indicate pairwise comparisons. Although monkey 1 performed slightly better than monkey 2, they displayed similar patterns of performance across conditions. **c**, Distribution of reported colours and absolute angular error as a function of target colour in experiment 1 for each monkey for pro and retro trials. The distributions of reported colours for each condition and monkey were significantly non-uniform (entropy of report distribution significantly lower than entropy of the target distribution, all  $P < 0.001$ , bootstrap across  $n = 3,873$  (pro) and  $3,943$  (retro) trials for monkey 1 and  $n = 4,440$  and  $4,769$  trials for monkey 2). Details of this behaviour have previously been published<sup>30</sup>. **d**, Mixture model parameter fits of behaviour

pooled across monkeys for experiment 1 (bootstrap across  $n = 23$  sessions). **e**, Top, in a separate behavioural experiment (experiment 2), we fixed the total memory delay of the retro condition and systematically varied the length of the delay between stimulus offset and cue onset. Bottom, increasing the time before selection ( $x$ -axis) increased mean absolute angular error ( $53.1^\circ$ ,  $54.4^\circ$ , and  $57.8^\circ$  for 0.5 s, 1 s, and 1.5 s post-stimulus, respectively; distributions are 1,000 bootstrap resamples across  $n = 3,306$ ,  $3,287$ , and  $3,322$  trials, respectively). **f**, Mixture model parameter fits, pooled across monkeys (1,000 bootstrap resamples across  $n = 24$  sessions), for experiment 2. Linear regression showed that earlier cues improved the precision of memory reports in experiment 2 ( $\beta = 3.95 \pm 1.88$  (s.e.m.),  $P = 0.012$ , bootstrap) but did not significantly change the probability of forgetting (that is, random responses;  $\beta = 0.03 \pm 0.03$  (s.e.m.),  $P = 0.126$ , bootstrap). Bars and asterisks in all panels reflect two-sided uncorrected randomization tests:  $\cdot P < 0.1$ ,  $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ .



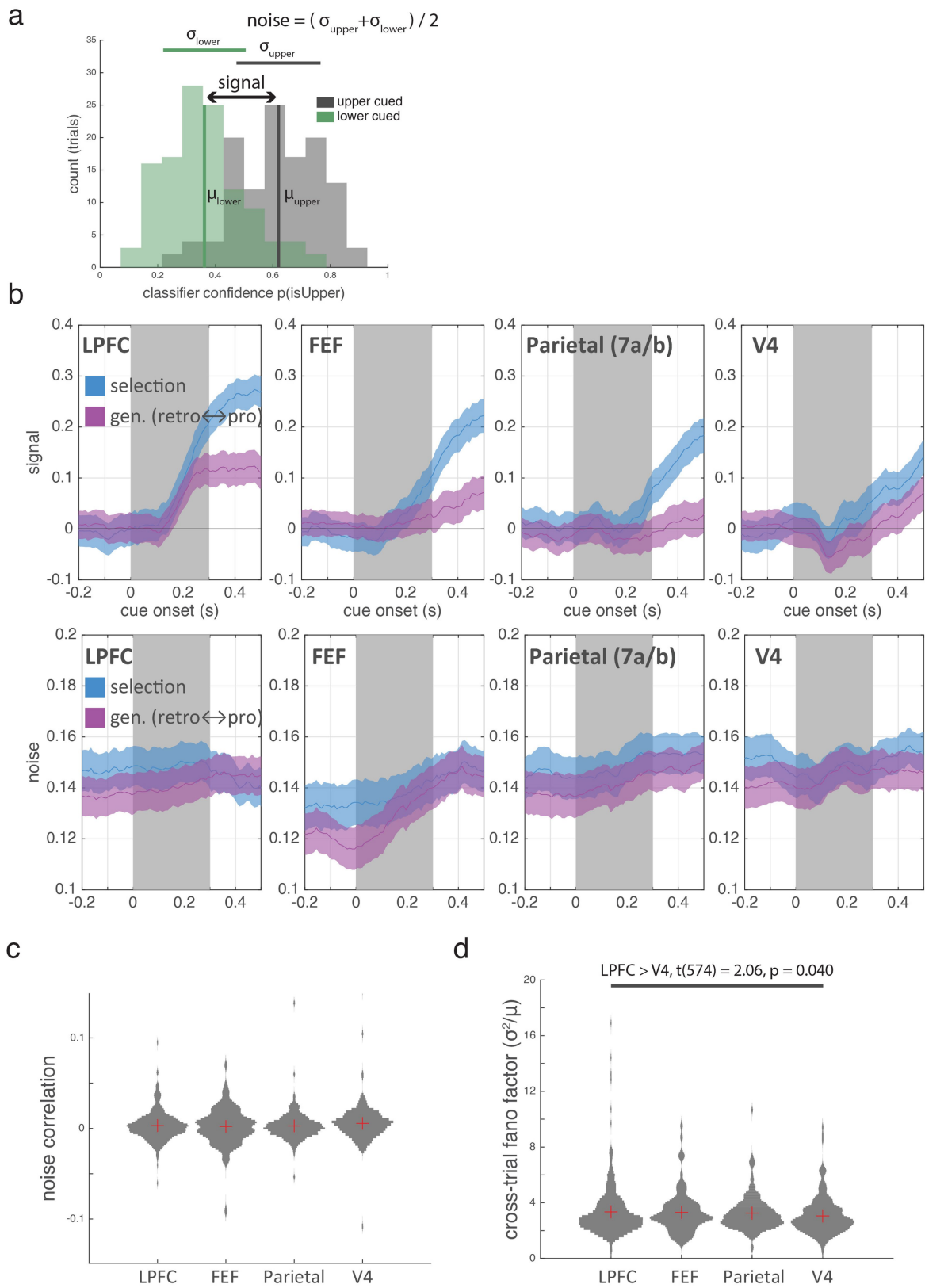
Extended Data Fig. 2 | See next page for caption.

# Article

**Extended Data Fig. 2 | Population size and neural responsiveness do not explain differences in classification performance across regions.** **a**, Firing rate of an example LPFC neuron around cue onset when the upper (grey) or lower (green) stimulus was cued in the retro (top) and pro (bottom) conditions. Shaded regions are s.e.m. across trials ( $n=161$  retro upper, 124 retro lower, 150 pro upper, and 121 pro lower trials). Insets, cues used for retro and pro trials. **b**, Percentage of neurons in each region of interest with firing rates that were significantly modulated by the selected location after cue onset on retro trials (trials pooled across cue sets 1 and 2). For each neuron, we quantified location selectivity using  $d'$  (Methods) and compared this value to a null distribution by permuting location labels across trials. All four regions showed strong selectivity: LPFC had 159 out of 590 neurons selective; FEF, 37 of 169; parietal, 49 of 301; V4, 62 of 318; all  $P < 0.001$  against chance of 5% (two-sided uncorrected binomial test). **c**, Mean classification accuracy (top, taken at 300 ms post-cue) and mean time to 55% classification accuracy (bottom) for the selection (left) and generalized (right) classifiers as a function of the number of neurons used for classification. This analysis controls for the total number of neurons recorded in each region. For each subpopulation of a specific size ( $x$ -axis), circles reflect average across 1,000 iterations using different randomly selected subpopulations of that size. Lines reflect best-fitting two-parameter power function (Methods). Error bars are 95% prediction intervals. For classifier accuracy (top row):  $n = 35, 10, 19,$  and  $22$  unique population sizes for LPFC, FEF, parietal and V4, respectively. For classifier timing (bottom left and right):  $n = 35/32, 10/8, 19/4,$  and  $21/20$  for selection/generalization in LPFC, FEF, parietal and V4, respectively. The reduction in the number of data points in the bottom plots reflects the fact that, for some neuron counts, classifiers never reached 55% classification accuracy on any iteration. Asterisks indicate significance of projected classification for a given region compared to the measured classification in LPFC at the maximum number of neurons (two-sided  $z$ -test, not corrected for multiple comparisons). Selection classification accuracy: FEF  $P = 2.18 \times 10^{-10}$ ; parietal  $P = 1 \times 10^{-16}$ ; V4  $P < 1 \times 10^{-16}$ . Generalization classification accuracy: FEF  $P < 1 \times 10^{-16}$ ; parietal  $P < 1 \times 10^{-16}$ ; V4  $P < 1 \times 10^{-16}$ . Selection classification timing: FEF  $P = 0.054$ ; parietal  $P = 1.02 \times 10^{-4}$ ; V4  $P < 6.94 \times 10^{-8}$ . Generalization classification timing: FEF  $P = 0.203$ ; parietal  $P = 1.11 \times 10^{-13}$ ; V4  $P < 1 \times 10^{-16}$ . **d**, Neuron dropping curves as in **c**, except analysis was restricted to neurons with a significant evoked response to cue onset to control for potential differences in responsiveness across regions (Methods). For classifier accuracy (top row):  $n = 23, 5, 8,$  and  $8$  unique population sizes for LPFC, FEF, parietal and V4, respectively. For classifier timing (bottom left and right):  $n = 23/22, 5/4, 8/0,$  and  $8/8$  for selection/generalization in LPFC, FEF, parietal and V4, respectively. Selection classification accuracy: FEF  $P < 1 \times 10^{-16}$ ; parietal  $P < 1 \times 10^{-16}$ ; V4  $P < 1 \times 10^{-16}$ . Selection classification timing: FEF  $P < 1 \times 10^{-16}$ ; parietal  $P < 1 \times 10^{-16}$ ; V4

$P < 1 \times 10^{-16}$ . Generalization classification accuracy: FEF  $P = 0.001$ ; parietal  $P = 0.021$ ; V4  $P = 0.002$ . Generalization classification timing: FEF  $P < 1 \times 10^{-16}$ ; V4  $P < 1 \times 10^{-16}$ . **e**, To determine whether there were sub-populations of selective neurons in a region with greater selectivity than the overall population, we ranked neurons in each region by their ability to support the selection (left) or generalized (right) classifier (Methods). Neurons with firing rates that yielded large magnitude (and sign consistent)  $d'$  values for the cued location (upper or lower) across both retro cue sets will support selection classifier performance (left). We quantified this by projecting these two  $d'$  values onto the identity (red lines) and taking the absolute value of the resulting vector. Neuron 1 is ranked higher than neuron 2 because of its larger magnitude projection onto the identity. A similar procedure can be used to rank neurons for generalization from pro to retro trials (right) by repeating the procedure on the basis of selectivity for 'pro cue set 1' and 'retro cue set 2'. **f**, Neuron dropping curves (as in **c**), except that neurons are added to the analysis on the basis of their selectivity or generalization, as described in **d**. Shaded region is 95% confidence intervals of best linear fit (which fit better than power functions) (Methods). Even when selecting ideal subpopulations from each region, no region significantly exceeded LPFC performance. Performance now decreases as  $n$  increases because, owing to our ranking procedure, later cells are by design less able to support performance on withheld cues (whether within selection or across selection or attention). These later neurons may still be weighted heavily by the classifier (owing to good performance on the training set) and so negatively affect performance at test. This is exemplified by the projections onto one axis, as indicated by the vertical dashed lines in **d**, showing a greater weighting for neuron 2, despite it not facilitating generalization. For classifier accuracy (top row):  $n = 35, 10, 19,$  and  $22$  unique population sizes for LPFC, FEF, parietal and V4, respectively. For classifier timing (bottom left and right):  $n = 35/35, 10/10, 19/18,$  and  $22/22$  for selection/generalization in LPFC, FEF, parietal and V4, respectively. **g**, To examine 'bottom-up' information flow about low-level sensory aspects of the cue, we trained classifiers to discriminate the variants of each cue, using cross-validation across subsets of trials (Methods). **h**, Neuron dropping curves (as in **c**) for these 'cue appearance' classifiers. Cue appearance classifiers yielded a qualitatively different pattern of performance, with V4 showing superior classification performance at cue offset (left) and faster classification onset (right). Asterisks indicate significance of projected classification for a given region compared to the measured classification in LPFC at the maximum number of neurons (two-sided  $z$ -test, not corrected for multiple comparisons).  $n = 35, 10, 19,$  and  $22$  unique population sizes for LPFC, FEF, parietal and V4, respectively. Classification accuracy: FEF  $P = 0.282$ , parietal  $P < 1 \times 10^{-16}$ , V4  $P < 1 \times 10^{-16}$ . Classification timing: FEF  $P = 0.005$ , parietal  $P = 4.24 \times 10^{-16}$ , V4  $P = 2.27 \times 10^{-8}$ .  $\cdot P < 0.1$ ,  $* P < 0.05$ ,  $** P < 0.01$ ,  $*** P < 0.001$ .



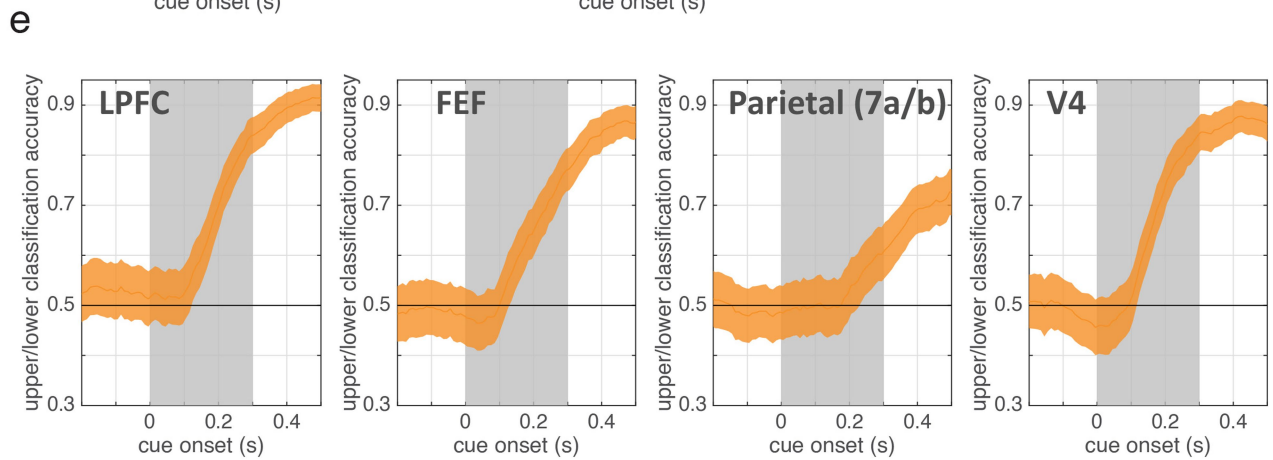
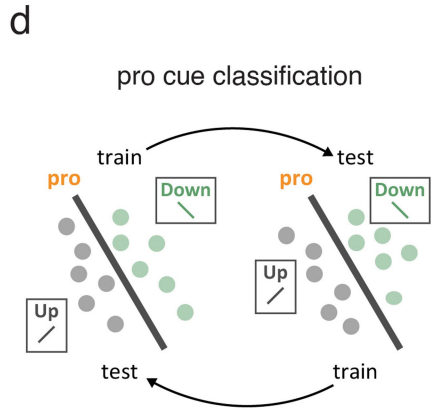
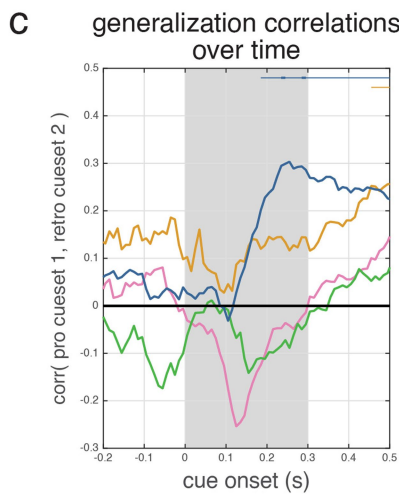
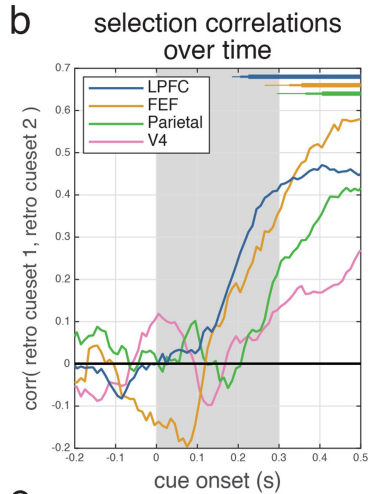
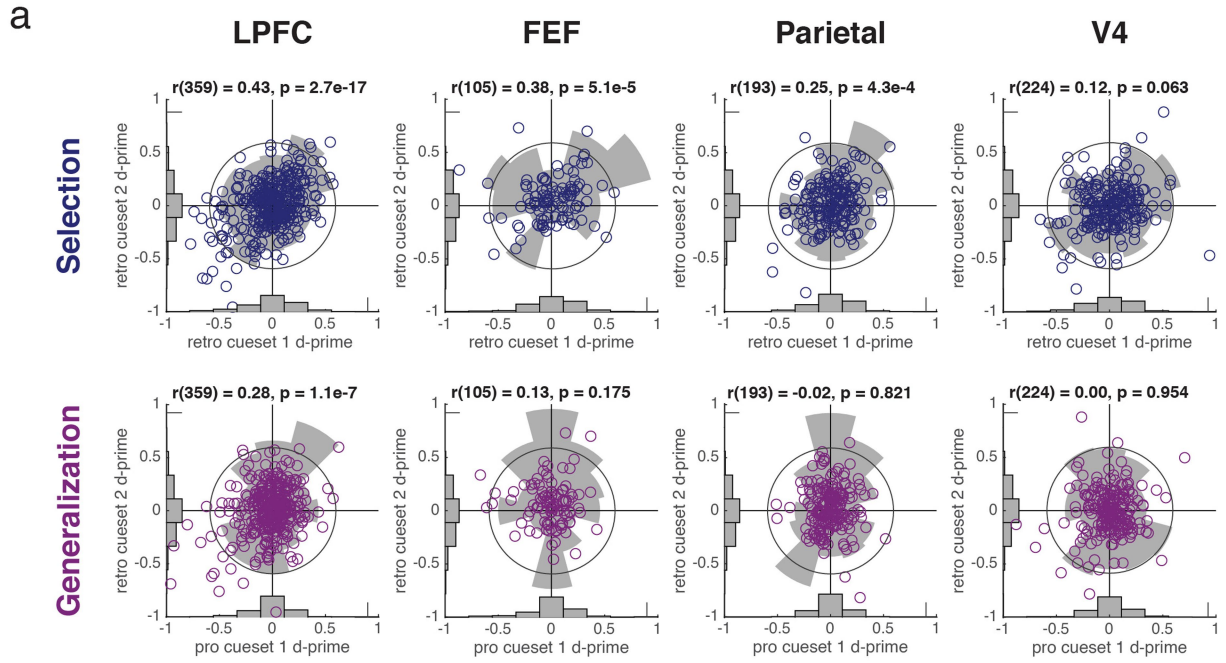


Extended Data Fig. 3 | See next page for caption.

# Article

**Extended Data Fig. 3 | Successful classification was driven by increases in signal.** **a**, Example histogram of classifier confidence across 'upper cued' and 'lower cued' trials for the LPFC selection classifier in the 500 ms after cue onset. Classifier confidence measures the distance of neural activity from the hyperplane identified by the classifier. Signal is the difference between the means of the two trial distributions; noise is their average s.d. **b**, For both the 'selection' and 'generalization' classifiers, signal (top row) tracks classification performance (Fig. 2) much better than noise (bottom row), suggesting that classifier performance was due to an increase in signal and not a decrease in noise. Shading shows s.e.m. Distribution estimated from 1,000 iterations of

classifiers trained and tested on random samples of  $n = 60$  trials (Methods). **c**, Mean noise correlation among neurons entering the 'selection' and 'generalization' analyses described in Fig. 2. Noise correlations were based on mean firing rates over the interval from 0 to 500 ms after the cue. There were no significant differences between regions. **d**, Fano factor ( $\sigma^2/\mu$ ) of single-neuron firing rates across trials (averaged from 0 to 500 ms after the cue). The ratio was significantly larger in LPFC than V4 but no other comparisons were significant (horizontal bar; two-sided uncorrected  $t$ -test). **c**, **d**, Violin plots show distribution of values based on 1,000 bootstrapped resamples of  $n = 60$  trials (Methods). Red crosses indicate mean.

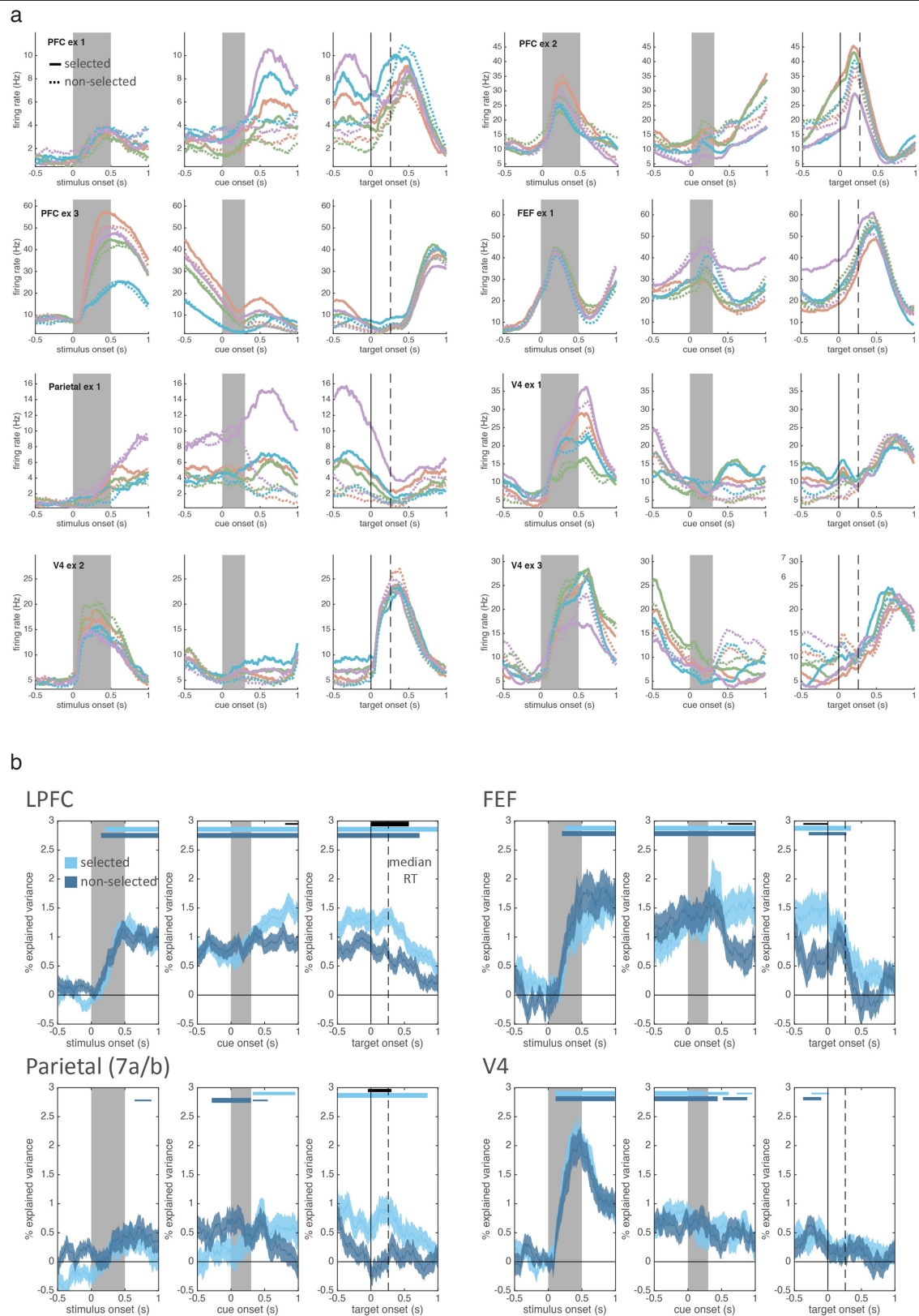


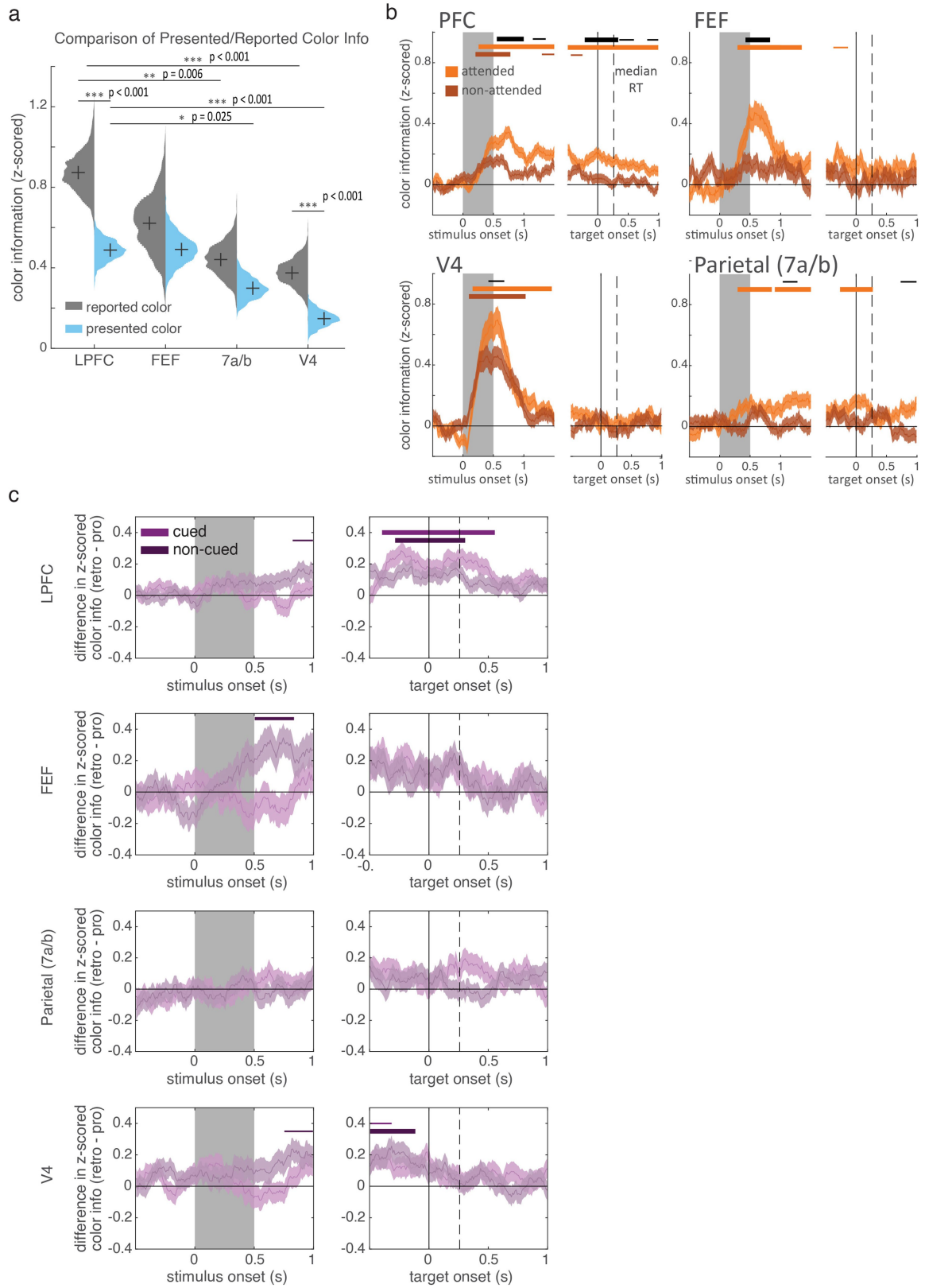
**Extended Data Fig. 4** | See next page for caption.

# Article

**Extended Data Fig. 4 | Neural responses in prefrontal cortex were similar across cue sets and tasks.** **a**, Distribution of selectivity across neurons for the selected location (top row) and for the selected and attended location (bottom row). Selectivity was taken as the normalized difference in firing rate ( $d'$ ) between 'upper' and 'lower' trials evoked by the two retro cue sets (top) and by pro cue set 1 and retro cue set 2 (bottom) (Methods). Firing rate was computed at the end of the cue period (300 ms after cue onset). Positive  $d'$  values indicated that the neuron was more active when the upper sample was cued. Rose plots in the background show the histogram of neurons binned by angle (grey circle indicates scale; density = 0.1). Bar plots along axes show histogram of marginal distributions (grey ticks on axes indicate scale; density = 0.2). Statistical tests are Pearson's  $r$ . **b**, Selection correlation values (as in **a**) computed over time around cue onset. Bars along top indicate correlations greater than zero:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick lines,

respectively (one-sided uncorrected bootstrap;  $n = 1,000$  resamples of trials). **c**, Generalization correlation values computed over time around cue onset, as in **b**. **d**, Schematic of classifier trained to discriminate the neural response to two cue conditions on pro trials. Performance was calculated as the cross-validated classification accuracy (tenfold cross-validation on each of 1,000 random resamples of trials) (Methods). **e**, Mean  $\pm$  s.e.m. classification accuracy of the pro cues, relative to cue onset, for all four brain regions. Distribution was defined across 1,000 random resamples of trials. This analysis captures a mixture of information about the control of attention (up or down) and information about the visual appearance of the cue itself. These results show that these two conditions are separable in all brain regions, and so any failure in cross-classification performance (Fig. 2d, purple traces) is not due to poor separability of the attention conditions.

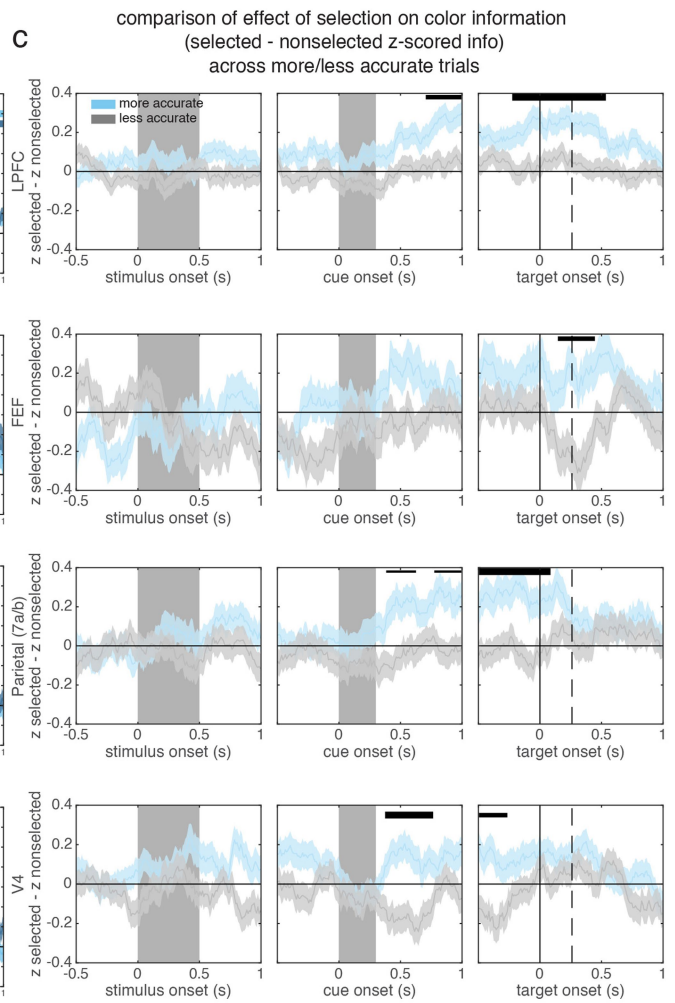
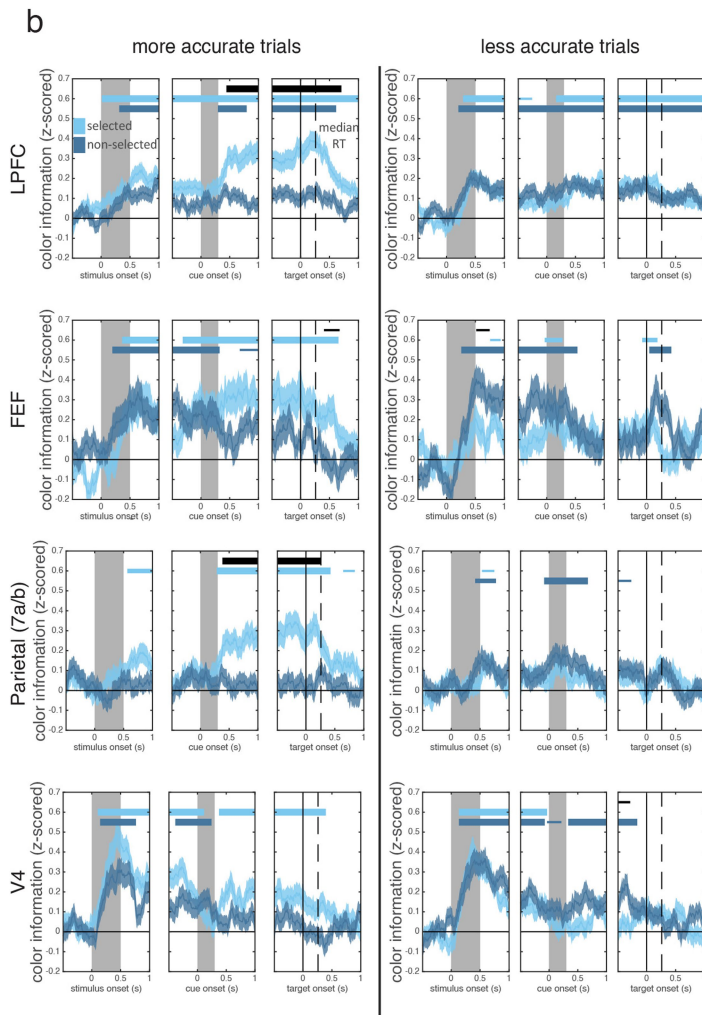
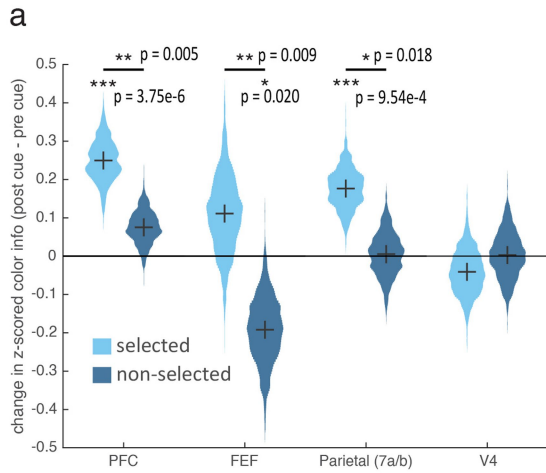




Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Comparison of information about the reported/presented colour, the attended/unattended item, and the memory of items on prospective and retrospective trials. a,** Mean z-scored colour information for the reported colour (grey) and the colour of the presented, selected, item (light blue). Information was calculated on firing rates in a 200-ms window before onset of the response colour wheel for all neurons. Distributions show bootstrapped estimates of the mean across neurons (LPFC: 570 neurons, FEF: 163 neurons, parietal: 292 neurons, V4: 311 neurons). Horizontal lines indicate pairwise comparisons. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-sided uncorrected randomization tests). **b,** Mean  $\pm$  s.e.m. z-scored colour information for the attended and non-attended colour on pro trials. LPFC: 543 neurons, FEF: 160 neurons, parietal: 272 neurons, V4: 300 neurons. Horizontal bars indicate significant information for the attended

item (light orange), the non-attended item (dark orange), and significant differences in information about the attended and non-attended items (black). Bar width indicates significance:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick, respectively (two-sided cluster-corrected  $t$ -tests). **c,** Mean  $\pm$  s.e.m. difference in z-scored colour information between retro and pro trials for the cued item (selected – attended; light purple) and uncued item (non-selected – non-attended; dark purple). Positive values reflect more information about an item on retro trials. LPFC: 511 neurons, FEF: 146 neurons, parietal: 258 neurons, V4: 285 neurons. Horizontal bars indicate significant differences from zero (that is, differences between retro and pro) for the cued item (light purple) and the non-cued item (dark purple). Bar width indicates significance:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick, respectively (two-sided cluster-corrected  $t$ -tests).

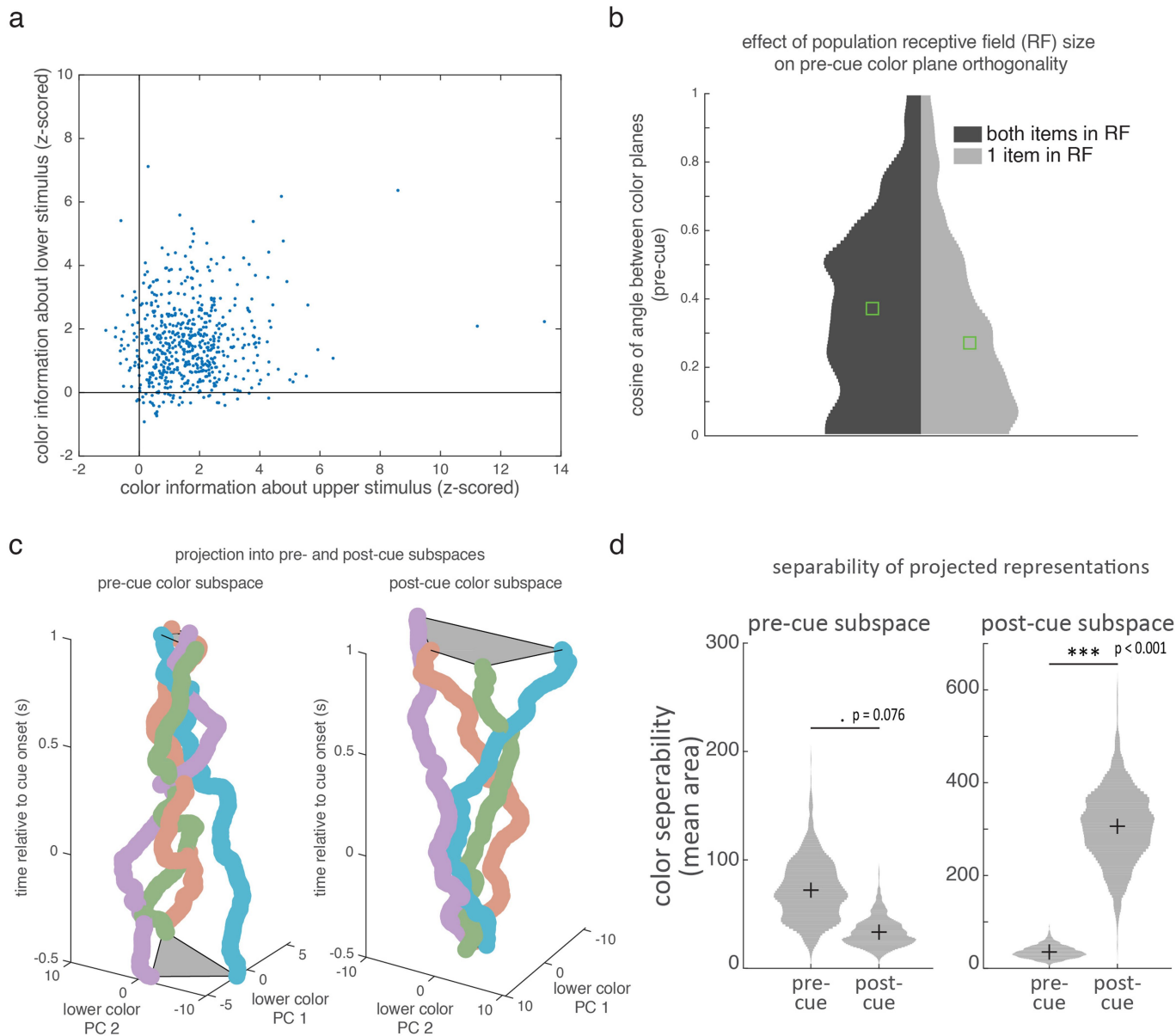


Extended Data Fig. 7 | See next page for caption.



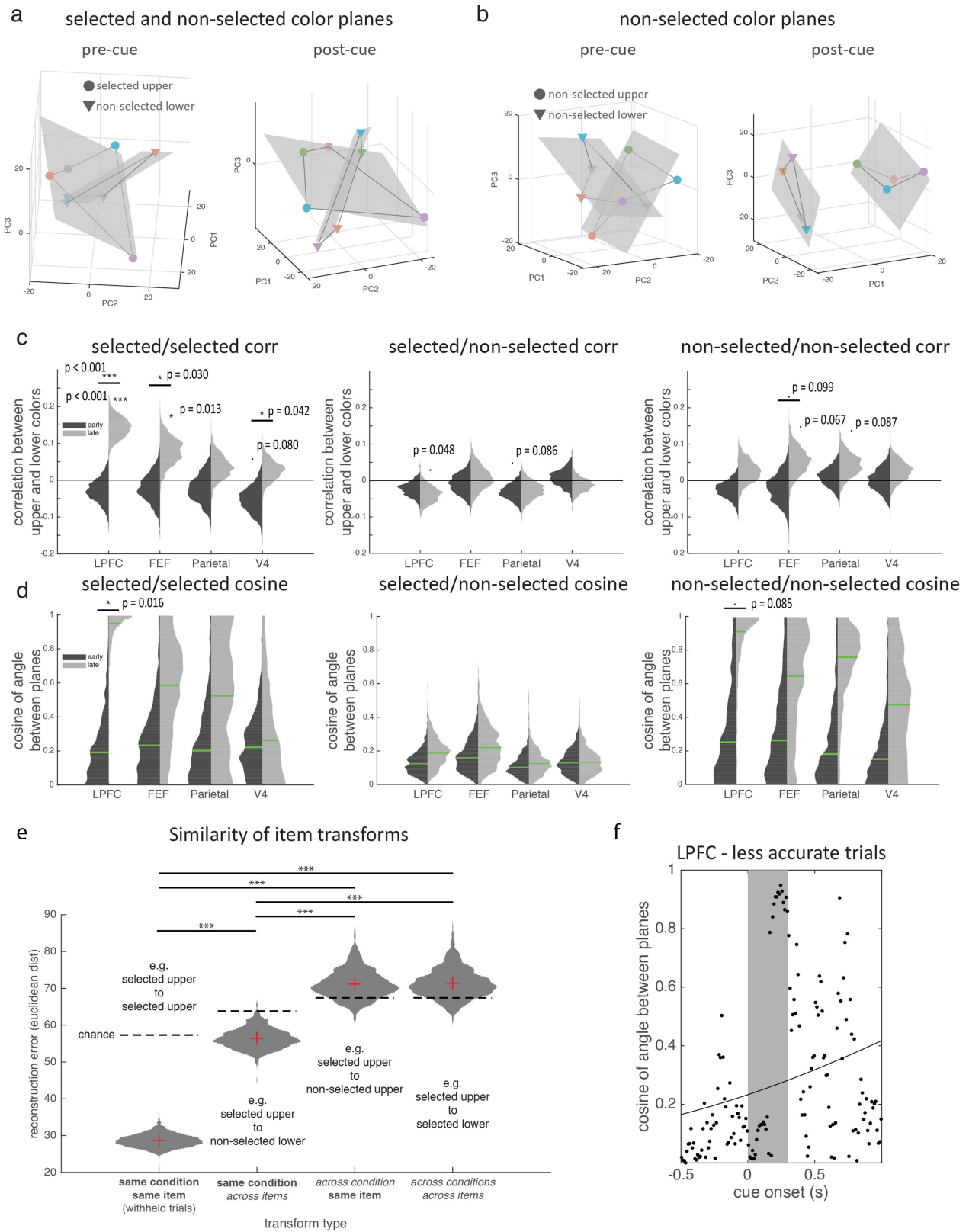
**Extended Data Fig. 7 | The effect of selection on colour information was greater when memories were more accurate.** **a**, Selection enhanced the representation of the selected item in frontal and parietal regions and reduced the representation of the unselected item in FEF. The y-axis shows the increase in colour information after selection (post-cue period: 200 to 500 ms after cue offset), relative to information before selection (pre-cue period: -300 to 0 ms before cue onset). Violin plots show the distribution of this difference, estimated by 1,000 bootstrapped resamples of neurons (LPFC: 577 neurons, FEF: 170 neurons, parietal: 299 neurons, V4: 316 neurons). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-sided uncorrected paired  $t$ -tests). **b**, Mean  $\pm$  s.e.m. z-scored colour information for the selected (light blue) and non-selected item (dark blue) on retro trials, for trials with more accurate behavioural responses (left; error was less than median error) and less accurate behavioural responses (right; error was greater than median error). LPFC: 457/472 neurons, FEF:

134/135 neurons, parietal: 235/241 neurons, V4: 248/267 neurons for left/right, respectively. Plots follow Fig. 3. Horizontal bars indicate significant information for the selected item (light blue), the non-selected item (dark blue), and significant differences in information about the selected and non-selected items (black). Bar widths indicate significance:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick, respectively (two-sided cluster-corrected  $t$ -tests). **c**, Mean  $\pm$  s.e.m. difference in z-scored colour information about the selected and non-selected items for more accurate and less accurate trials. LPFC: 435 neurons, FEF: 125 neurons, parietal: 221 neurons, V4: 240 neurons. As in **b**, trials were split on the basis of angular error (relative to median error). Positive values reflect more information about the selected item than the non-selected item. Horizontal bars indicate significant differences between more and less accurate trials; width indicates significance:  $P < 0.05$ , 0.01, and 0.001 for thin, medium, and thick, respectively (two-sided cluster-corrected  $t$ -tests).



**Extended Data Fig. 8 | Distributed representations of colour in prefrontal cortex were transformed over time. a**, Mean z-scored colour information for the upper (x-axis) and lower (y-axis) stimuli immediately before selection cue onset (average over  $-500$  to  $0$  ms before the selection cue) for LPFC (583 neurons). Most neurons carried some amount of information about both items (that is, neurons did not lie along the axes). **b**, To check whether neurons that primarily carried information about just one item were driving the orthogonality between the colour planes in LPFC before the selection cue, we re-computed the cosine of the angle between the best-fitting planes using populations of neurons with significant colour information about one item only or both items (see Methods for description of this test). Histograms show the distribution of the cosine of the angle between the best-fitting planes for the upper and lower stimuli during the pre-cue period for these ‘both’ and ‘1 item’ populations of neurons (with each population subsampled to an equal number of neurons) (Methods). Distributions were estimated from 1,000 resamples of trials. Green squares indicate median values. While the ‘both’ neurons did display slightly less orthogonality than the ‘1 item’ neurons, this difference was not significant ( $P > 0.4$ , two-sided bootstrap of difference).

Cosine angles are not zero for ‘1 item’ neurons because ‘1 item’ neurons still contain subthreshold information ( $P > 0.05$ ) about the other item, as seen in **a**, and subsampling cells in this way decreases statistical power, thereby inflating low cosine values. **c**, Population trajectories for lower colours, over time, as projected into the lower colour subspace defined either before or after selection (left and right, respectively). Follows Fig. 4e. The lower colour subspace was defined as a 2D space that maximally explained variance across the four lower colours (Methods). As for the upper colour (Fig. 4e), temporal cross-generalization was poor, suggesting that the colour information was represented in different subspaces before and after the selection cue. **d**, Before selection, colour representations in LPFC are better separated using the pre-selection subspace. After selection, colours are better separated in the post-selection subspace. Separability was measured as the area of the quadrilateral defined by the responses to colours (**c**, Fig. 4e), projected into either the pre-selection or post-selection subspaces (left and right columns in each plot; area averaged across upper and lower items). Violin plots show distributions estimated from 1,000 resamples of trials. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-sided bootstrap of difference).



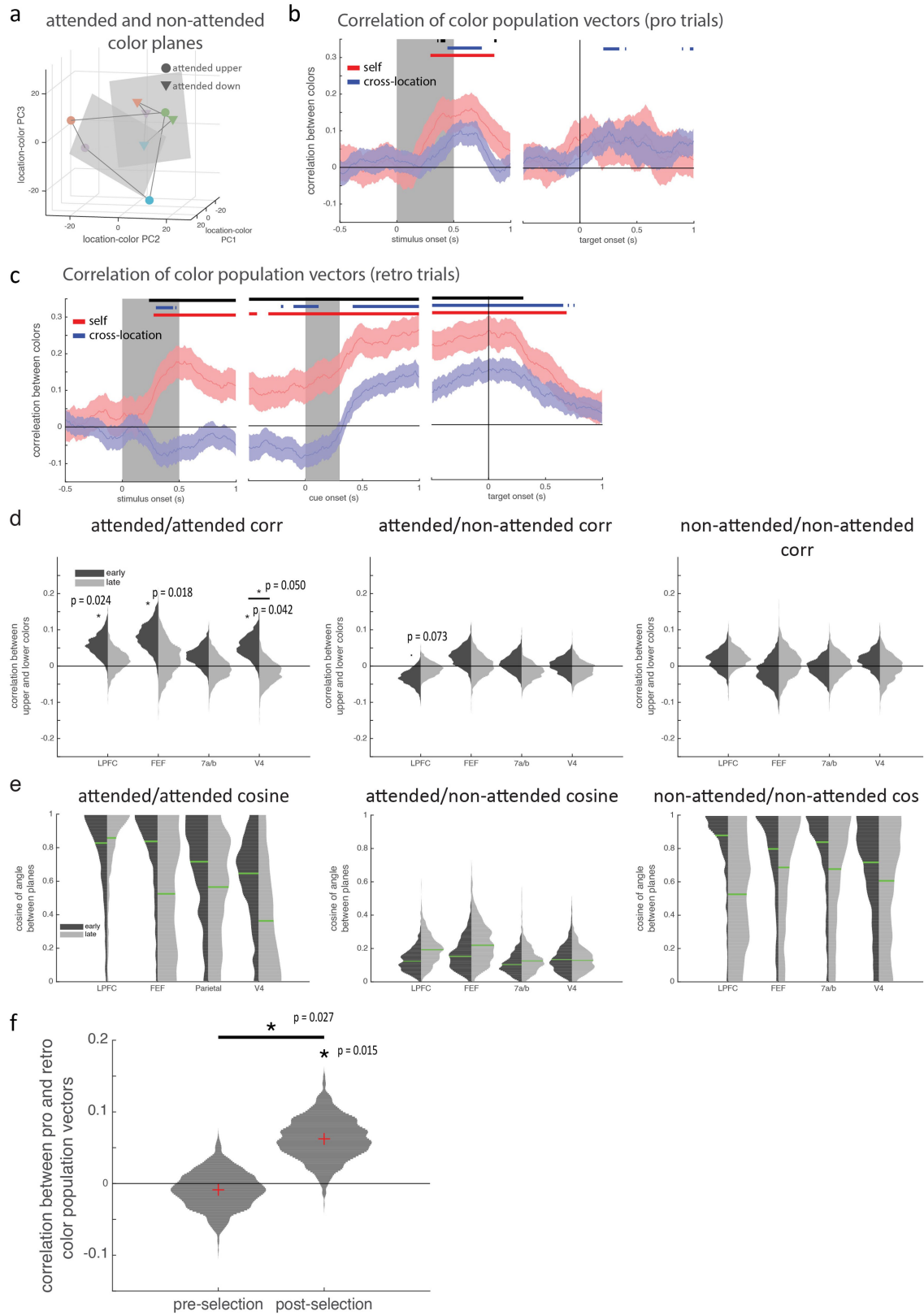
Extended Data Fig. 9 | See next page for caption.

# Article

## **Extended Data Fig. 9 | The alignment of selected items was greater in prefrontal cortex than other brain regions, was greater than the alignment of non-selected items, and was greater when memories were more accurate.**

**a.** Projected population responses for selected upper and non-selected lower colours, computed as in Fig. 4a. The selected and non-selected colours remain orthogonal after the selection cue (main text). **b.** Projected population responses for non-selected upper and non-selected lower colours. As with the selected colour planes, the non-selected colour planes appear parallel after the selection cue. **c.** Mean correlation between the population representation of each colour in the upper and lower position during retro trials, when both items were selected (left), one item was selected and another item was non-selected (middle), and when both items were non-selected (right). Correlation was measured during an 'early' time period during the delay (dark grey; 150–350 ms after the offset of the stimulus) and a 'late' time period during the delay (light grey; 200–0 ms before the onset of the colour wheel). Correlation was measured after subtracting the mean response at each location (Methods). Violin plots show bootstrapped distributions estimated from 1,000 resamples of trials. Horizontal lines indicate pairwise comparisons (two-sided uncorrected bootstrap of difference). Lone asterisks denote two-sided uncorrected bootstrap versus zero: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . **d.** Cosine of the angle between the best-fitting planes for the upper and lower stimuli. Planes were fit to selected and non-selected items during both the early and late time periods (as in c). Histograms show full distribution, estimated from 1,000 resamples of trials; green lines indicate median values. Horizontal lines indicate pairwise comparisons: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-sided uncorrected bootstrap of difference). **e.** To find out

whether the selection process transformed the cued and non-cued item in similar ways, we estimated the transformation matrices that mapped pre-cue representations of an item onto their post-cue representation (Methods, Supplementary Discussion 3). Then, we tested whether these transformations were able to reconstruct representations on withheld trials. Transformations were tested on the same condition (withheld trials; first column); on the other item in a condition (for example, applying the transformation of a selected upper item to a non-selected lower item; second column); on the same item, but in a different condition (for example, applying the transformation of a selected upper item to a non-selected upper item; third column); and on the other item in a different condition (for example, applying the transformation of a selected upper item to a selected lower item; fourth column). Violin plots show distributions of these mean reconstruction errors estimated from 1,000 resamples of trials. Red crosses indicate the distribution mean, dashed lines show reconstruction error expected by chance (estimated by random shuffle) (Methods). The results indicate a common component to the transformation of the selected and non-selected item in the same condition (second column) but there was also an item-specific transformation (reflected in the lower reconstruction error for the same item; first column). Horizontal lines show pairwise comparisons: \*\*\* $P < 0.001$  by two-sided uncorrected bootstrap of difference. **f.** The selected upper and selected lower colour planes do not align on inaccurate trials. Figure follows Fig. 4b, but shows data for trials in which absolute angular error was greater than the median error. Black markers show the cosine of the angle ( $y$ -axis) between the two colour planes around the time of cue onset ( $x$ -axis) and black line shows the best-fitting logistic function.



**Extended Data Fig. 10** | See next page for caption.

# Article

## Extended Data Fig. 10 | Colour representations of the attended item were immediately aligned on prospective trials. **a**, Population responses 200 ms after stimulus offset on pro trials (projected into a reduced subspace for visualization). As in Fig. 4a, markers indicate mean position of population activity for each condition (binned by the colour and location of the attended item) in a subspace spanned by the first three principal components that explain the most variance across all eight conditions. **b**, Mean $\pm$ s.e.m. correlation of population vectors representing colours at the same location (self; red line) or between locations (cross-location; blue line) on pro trials. Correlations were measured after subtracting the mean vector at each location (as in Fig. 4c; Methods). Distribution was estimated from 1,000 resamples of trials. Self-correlation was computed on held-out trials and provides an upper-bound on the between-location correlation values, given the noise level. Bars reflect uncorrected two-sided bootstrap ( $P < 0.05$ ) for each correlation type against zero (red and blue) and between each other (black). **c**, As in **b**, but for retro trials. **d**, Mean correlation between the population representations of each colour in the upper and lower position during pro trials, when both items were attended (left), one item was attended and another item was non-attended (middle), and when both items were non-attended (right). Correlation was measured during an 'early' time period during the delay (dark grey; 150–350 ms after the offset of the stimulus) and a 'late' time period during the delay (light grey; 200–0 ms before the onset of the colour wheel). Correlation

was measured after subtracting the mean response at each location (Methods). Violin plots show distributions, estimated from 1,000 resamples of trials. Horizontal lines indicate pairwise comparisons (two-sided uncorrected bootstrap of difference) and lone asterisks reflect two-sided uncorrected bootstrap against zero: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . **e**, Cosine of the angle between the best-fitting planes for the upper and lower stimuli. Planes were fit to attended and non-attended items during both the early and late time periods, as in **d**. Histograms show full distribution, estimated from 1,000 resamples of trials; green lines indicate median values. **f**, Mean correlation between the population representation for each colour during pro trials and the representations during the early or late time periods of retro trials. Correlation was computed between the colour representations taken from the 300 ms before the onset of the response wheel on pro trials and the colour representations taken from either a pre-selection period (left distribution; –300 to 0 ms before cue) or a post-selection period (right distribution; –300 to 0 ms before response wheel onset) on retro trials. Correlations were measured after subtracting the mean vector at each location, as in Fig. 4c (Methods). Violin plots reflect the distribution, estimated from 1,000 resamples of trials. Horizontal line indicates pairwise comparison (two-sided uncorrected bootstrap of difference) and lone asterisks reflect two-sided bootstrap against zero: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

was measured after subtracting the mean response at each location (Methods). Violin plots show distributions, estimated from 1,000 resamples of trials. Horizontal lines indicate pairwise comparisons (two-sided uncorrected bootstrap of difference) and lone asterisks reflect two-sided uncorrected bootstrap against zero: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . **e**, Cosine of the angle between the best-fitting planes for the upper and lower stimuli. Planes were fit to attended and non-attended items during both the early and late time periods, as in **d**. Histograms show full distribution, estimated from 1,000 resamples of trials; green lines indicate median values. **f**, Mean correlation between the population representation for each colour during pro trials and the representations during the early or late time periods of retro trials. Correlation was computed between the colour representations taken from the 300 ms before the onset of the response wheel on pro trials and the colour representations taken from either a pre-selection period (left distribution; –300 to 0 ms before cue) or a post-selection period (right distribution; –300 to 0 ms before response wheel onset) on retro trials. Correlations were measured after subtracting the mean vector at each location, as in Fig. 4c (Methods). Violin plots reflect the distribution, estimated from 1,000 resamples of trials. Horizontal line indicates pairwise comparison (two-sided uncorrected bootstrap of difference) and lone asterisks reflect two-sided bootstrap against zero: \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

Corresponding author(s): Tim BuschmanLast updated by author(s): 2021/01/05

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection	Stimuli were presented and behavioral responses registered using Psychtoolbox (version 3) and Matlab (version 2015a). Behavioral training code is available upon request. Electrophysiological data was collected using an amplifier from Blackrock Microsystems (Cerebus software, version 7). Eye tracking data was collected using an EyeLink 1000 (SR Research, software version 5.09). Neuron waveforms were sorted using Plexon's Offline Sorter (version 4).
Data analysis	Data were analyzed using built-in functions and custom code written in Matlab. Built-in functions are noted in the associated methods section (e.g., <code>pca</code> for dimensionality reduction; <code>fitlinear</code> for logistic regression). Equations for non-standard statistics are provided in the methods. Code for custom functions are provided in the manuscript as a reference to the original source or are included in an online public repository ( <a href="https://github.com/buschman-lab/SelectionFromWorkingMemory">https://github.com/buschman-lab/SelectionFromWorkingMemory</a> ). Full code is available upon reasonable request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data supporting all figures are included with the manuscript. Raw electrophysiological and behavioral data are available from the corresponding author upon reasonable request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As detailed in the manuscript, a total of 1,541 neurons were recorded from 4 brain regions across 2 animal subjects. The number of subjects (2) and the number of neurons recorded per region (~180-680) follows previous work using similar approaches (e.g., Wessberg et al, 2000; Buschman and Miller, 2007; Mante et al, 2013; Siegel et al, 2015).
Data exclusions	Neurons were excluded from a particular analysis if they were recorded for fewer than a fixed number of trials in any condition of interest for that analysis (typically 10 trials). This exclusion criterion was established in advance and is described specifically for each analysis in the methods.
Replication	Independent experiments were performed in 2 animals, with 1,541 neurons recorded across 23 days. All data was included (except for exclusions noted above). There were no failed replication attempts (i.e., no animals failed to learn the task and no neural recordings were excluded).
Randomization	Each animal was exposed to every task manipulation. Within a session, task manipulations were randomized across trials. Neurons were recorded without bias, with electrodes placed to maximize signal-to-noise of the electrophysiological signal.
Blinding	All animals were assigned to a single experimental group, and so blinding was not necessary or possible. However, experimenters were blinded to experimental conditions during recording of neurons and during sorting of waveforms into single neurons.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

### Methods

n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Subjects were two 9 year old male rhesus macaques ( <i>Macaca mulatta</i> ). Animals were singly-housed in a large room that contained 2-4 other male macaques of similar ages.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field-collected samples were used in this study.
Ethics oversight	All experimental procedures were approved by the Princeton University Animal Care and Use Committee and were in accordance with the policies and procedures of the National Institutes of Health.

Note that full information on the approval of the study protocol must also be provided in the manuscript.