SEE CORRECTION FOR THIS ARTICLE
RESEARCH ARTICLE | NEUROSCIENCE

Neural signature of flexible coding in prefrontal cortex

Andrea Bocincova[a,1], Timothy J. Buschman[b,c], Mark G. Stokes[d], and Sanjay G. Manohar[a,d]

Edited by Robert Desimone, Massachusetts Institute of Technology, Cambridge, MA; received January 9, 2022; accepted August 15, 2022

The ability of prefrontal cortex to quickly encode novel associations is crucial for adaptive behavior and central to working memory. Fast Hebbian changes in synaptic strength permit forming new associations, but neuronal signatures of this have been elusive. We devised a trialwise index of pattern similarity to look for rapid changes in population codes. Based on a computational model of working memory, we hypothesized that synaptic strength—and consequently, the tuning of neurons—could change if features of a subsequent stimulus need to be "reassociated," i.e., if bindings between features need to be broken to encode the new item. As a result, identical stimuli might elicit different neural responses. As predicted, neural response similarity dropped following rebinding, but only in prefrontal cortex. The history-dependent changes were expressed on top of traditional, fixed selectivity and were not explainable by carryover of previous firing into the current trial or by neural adaptation.

synaptic plasticity | working memory | history-dependent neural selectivity | computational model of working memory | prefrontal cortex

Significance

Rapid changes in neural selectivity have been proposed as a potential mechanism for storing novel associations. Despite this potential mechanism's being well-recognized in computational models, direct neural evidence is still lacking. Here, we show that characteristic trial-to-trial changes in neural selectivity generated by a working memory model implementing fast Hebbian synaptic plasticity are also present in prefrontal cortex neural populations of monkeys performing a working memory task. Using a trialwise pattern similarity method to track these changes during the encoding of associations, we show that changes in neural selectivity followed the encoding of a new stimulus that breaks down an association between the features of a previous stimulus.

Adaptive behavior requires flexible cognition, but how flexibility is achieved in the brain remains an open question. Prefrontal cortex (PFC) has been one of the focal points in attempts to find the neural correlates of cognitive flexibility because certain cells in this area change their response adaptively depending on the current task (1). Working memory (WM) is the workspace of flexible cognition (2–5), where a common neural resource may be used to hold diverse contents which may rely on flexible neural coding (6–10). However, the nature of WM representations remains debated. Two broad classes of adaptive coding have been proposed. In stable-coding models, flexible neurons respond to many stimuli with complex, variable-gain responses but fixed selectivity (11), whereas in plastic models, neural selectivity for the same stimulus differs depending on the task or situation (12–14). Rapidly changing neural selectivity could be useful to bind novel combinations of features that have not previously been encountered, without the need to have preexisting conjunction-specific neurons.

How might neurons change their tuning from moment to moment? One way to achieve this is by quick modification of the efficacy of synapses. Synaptic changes such as short-term facilitation have been previously proposed to maintain information in WM (15–17). However, facilitation alone would not account for the formation of new associations (18). For this, plasticity must also be sensitive to the postsynaptic response. For example, fast Hebbian synaptic plasticity could support maintenance in WM (19, 20). Hebbian changes in synaptic strengths can form memory associations in the brain that are transient and, importantly, can be overwritten in a use-dependent manner (21). Accordingly, a number of computational models of WM implement Hebbian synaptic plasticity to simulate a range of cognitive memory effects (22–25). To date, little direct neural evidence has supported these rapidly associative neural codes, partly because methods have not yet been developed to detect and model them.

In these models, memory neurons receive inputs from a range of feature-coding neurons (Fig. 1*A*). According to the fixed-coding view, a pattern of sensory input will always activate the same pattern in the memory neurons (6, 26, 27). In contrast, if memory neurons undergo associative plasticity, then their selectivity could change each time a stimulus is encoded. Simulating this yields distinct predictions (24) (see Fig. 1*A* for model's architecture). Hebbian plasticity allows new combinations of coactivated sensory features to be encoded in synapses, providing flexibility in the coding of neurons. Critically, the pattern of activation in such flexible neurons depends not only on the stimulus but also on the prior state of the synaptic weights. According to Hebbian rules, synapses between neurons representing active features (f in Fig. 1*A*) and flexible conjunctive neurons (c in Fig. 1*A*) become stronger, while synapses connecting conjunctive neurons with inactive features erode. This facilitates maintenance of the active features on the current trial. However, these synapses are overwritten when the feature changes its association across trials, i.e., when a participant has to rebind an old feature into a new object. This prediction can be tested against neural data.

Author affiliations: [a]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, United Kingdom; [b]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540; [c]Department of Psychology, Princeton University, Princeton, NJ 08540; and [d]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom

Author contributions: A.B., M.G.S., and S.G.M. designed research; T.J.B. performed research; A.B., M.G.S., and S.G.M. analyzed data; and A.B., T.J.B., M.G.S., and S.G.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: andreabocincova@gmail.com.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2200400119/-/DCSupplemental.

Published September 26, 2022.

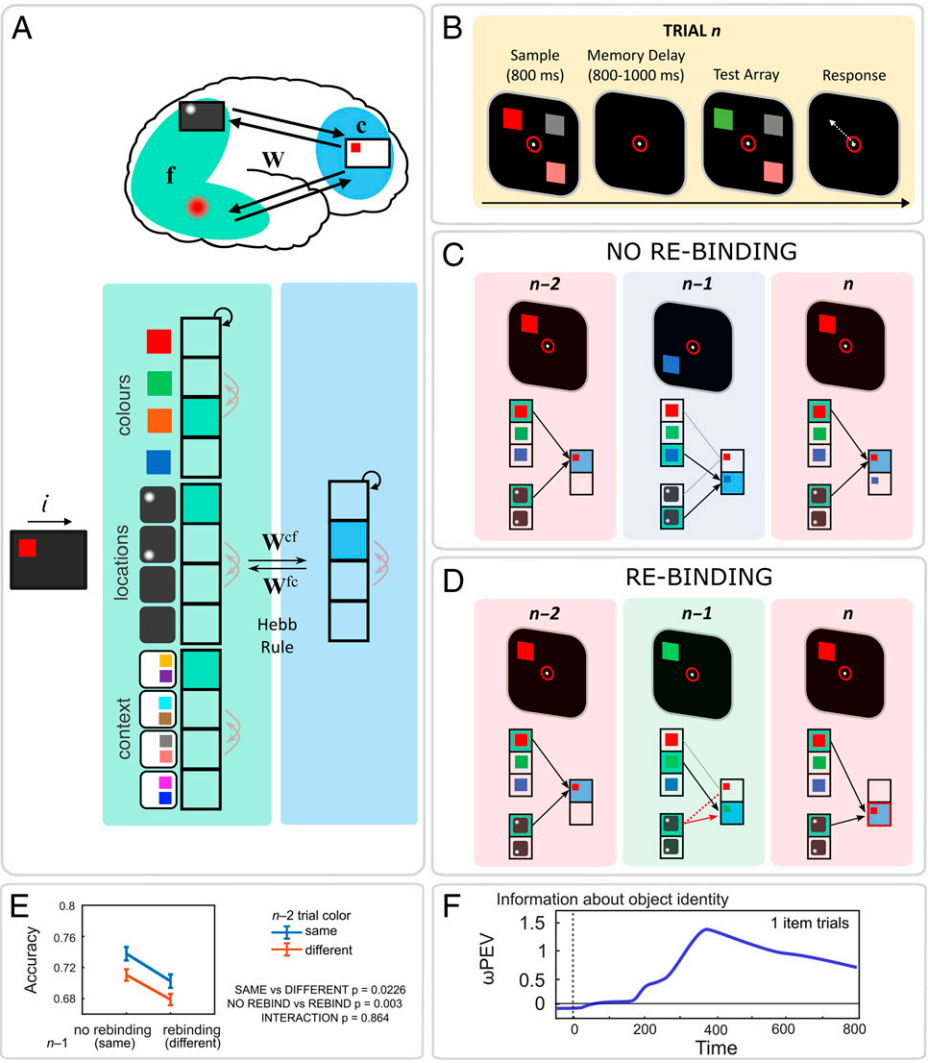

Fig. 1. Model architecture, trial sequence, and predictions based on associative plasticity. (*A*) The architecture of a neural model implementing fast Hebbian plasticity to encode new associations (24). The model is composed of two layers, a feature selective layer divided into three dimensions (color, location, and context) and a conjunction layer consisting of four freely conjunctive neurons. New feature associations are encoded as patterns of synaptic weights between active feature and conjunctive units. (*B*) Illustrative trial sequence from the change-localization task used by Buschman et al. (28). (*C*) Associative plasticity accounts predict that coding of an item will depend on trial history. If the feature binding in trial $n - 1$ does not directly interact with the synaptic memory of the item from $n - 2$, the synaptic trace stays intact and as a result coding remains the same. (*D*) However, if the item in trial $n - 1$ results in degradation of parts of the synaptic memory, as is the case when a feature changes its association, the conjunction unit that codes for the same object as in trial $n - 2$ is likely to change in trial n. This results in a drop in coding similarity between trial $n - 2$ and n. (*E*) Monkeys performed better when the trial n color at a location was the same as in trial $n - 2$, or the same as in trials $n - 1$ (i.e., no rebinding trials). (*F*) Neurons in lateral PFC encode the color of the item, as indexed by the proportion of variance explained by stimulus identity. Adapted from Buschman et al. (28).

Consider the task of remembering a color shown at a spatial location. As a result of synaptic changes, the same combination of features can be coded by a different pattern of neurons on different trials. For example, if the association between red and upper-left location is encoded by a conjunctive neuron on trial $n - 2$, then it will be encoded by the same neuron on trial n, as long as the association between the color and location is not broken in the intervening $n - 1$ trial. This is the case, for example, if trial $n - 1$ requires associating a new color with a different location (Fig. 1*C*), or if trial $n - 1$ has the same color–location pairing as $n - 2$. However, if a new color is associated with the same location in trial $n - 1$ (Fig. 1*D*), then the synaptic connection between the location and the originally active memory neuron is eroded as this new association is encoded via a different memory neuron. We term this "rebinding," because there is a new color–location pairing. When the original combination of features needs to be encoded again in trial n, the change in synaptic weights from trial $n - 1$ means that it will engage a different memory neuron (compared to trial $n - 2$). Thus, rapid associative plasticity predicts specific stimulus-driven changes in neural selectivity.

In this study we sought evidence for such changes in coding of neural populations in PFC as evidence for the presence of changes in synaptic efficacy. We analyzed neurophysiological data from two monkeys performing a WM change localization task for an array of colors (28) (see *Methods* and Fig. 1*B*). To ascertain whether disruption of previous associations shifted neural population coding, we looked for a drop in similarity of neural response between two trials where the same stimulus was encoded, using a novel trial-triplet pattern similarity analysis. We found that trial history affects coding in the following way: The pattern of neural activation elicited by a stimulus was less similar to the pattern elicited by the same stimulus if the trial separating these two trials involved recombination of its features, i.e., rebinding. This signature drop in similarity was consistent with predictions made by a computational model of

synaptic plasticity and was specific to the PFC. Moreover, this effect was expressed on top of traditional, fixed selectivity and was not explainable by carryover of previous firing into the current trial or by neural adaptation.

Results

A Mismatching Stimulus Changes the Population Code. Behavioral memory accuracy was greater both when the color at a location was identical on the previous trial and on the trial before that (Fig. 1*E*). To understand how neurons reflected associations at this timescale, we first quantified each neuron's activity in response to a stimulus and while it was maintained in WM. Across all trials, individual neurons encoded stimulus identity (Fig. 1*F*) and location (*SI Appendix*, Fig. S7). Triplets of trials were selected and classified based on whether the stimuli in trials n and $n-2$ were the same or different, and based on whether the features from the stimulus in trial $n-1$ had to be re-bound when encoding stimulus n (Fig. 2 *A* and *B*). This grouping was performed using one location at a time then

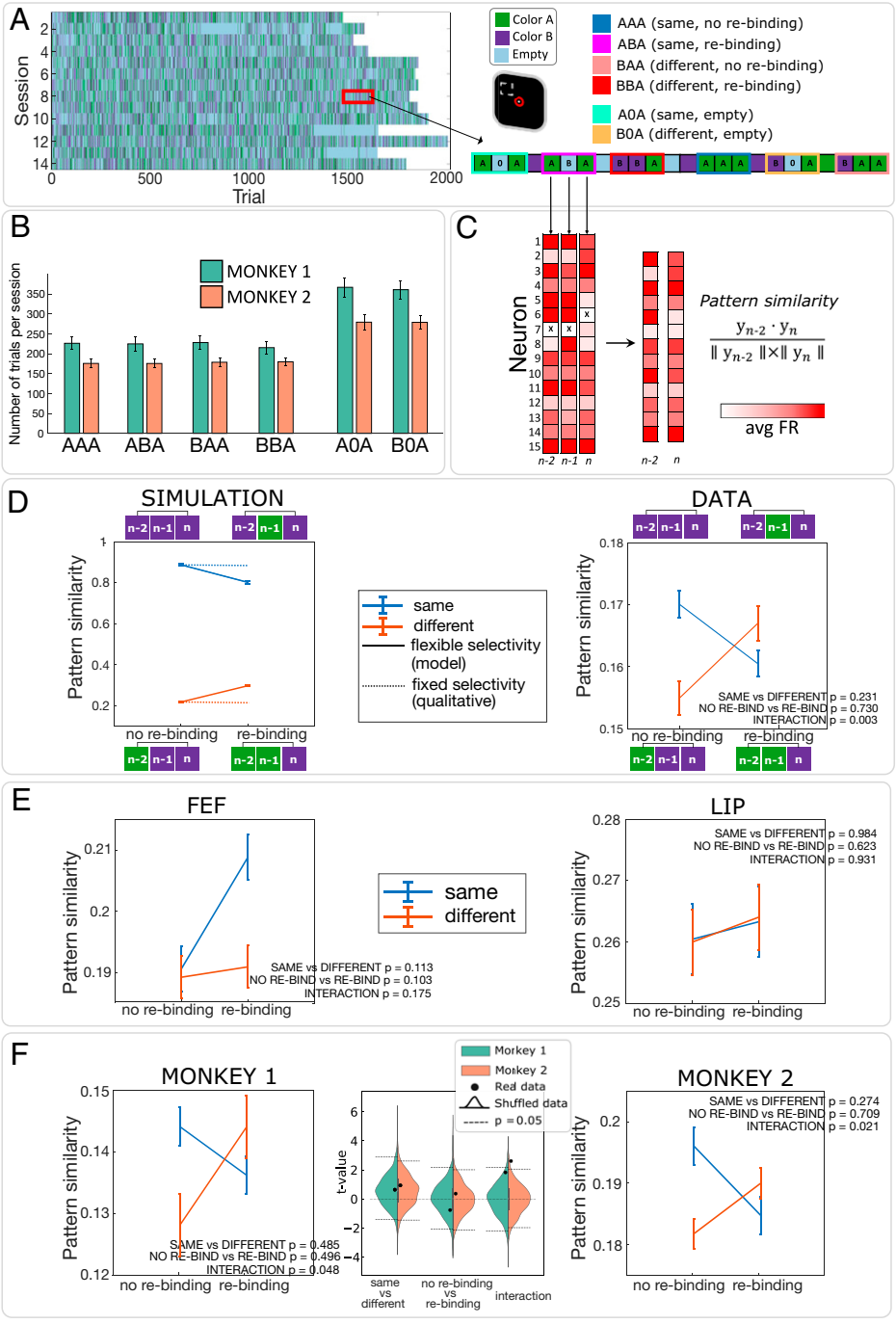

Fig. 2. Data analysis and results. (*A*) For each location (six locations total), one at a time, each trial was labeled based on the color presented in each trial (color A, color B, or empty). Six different triplet types were identified in the resulting trial sequence depending on whether the color in trial n and $n-2$ was the same or different and whether the color in trial $n-1$ was the same or different as in trial n or absent. (*B*) A histogram of the number of trials falling into the different triplet categories for each monkey. (*C*) Pattern similarity was calculated for each triplet by comparing the vector of activation (1x neurons, leaving out any neurons with missing data marked as "x") for trial n with trial $n-2$ using cosine similarity. (*D*) In line with the model's predictions (*Left*), the data showed a drop in similarity in the neural response to the same stimulus following a rebinding (*Right*). (*E*) This pattern was only observed in the PFC, not FEF or LIP. (*F*) Pattern similarity results plotted separately for each monkey (*Left* and *Right*) and the results of permutation analysis (*Middle*).

results were averaged across locations. We then compared the similarity of activity patterns across pairs of trials (Fig. 2*C*). With stable selectivity, pattern similarity should be high when the stimuli on those trials were the same but lower when they were different (Fig. 2*D*, *Left*).

In contrast, rapid Hebbian plasticity results in changes in selectivity when features are "re-bound," i.e., a feature is combined with different features. In this task, this occurs if a different color is reassociated with a location. This new binding should trigger a change in selectivity of flexible memory neurons (Fig. 1*D*). If a new color is bound to a location and then later the original color must be re-bound (ABA trial), this results in a drop in pattern similarity between the two A trials, which code the same stimulus preceding and following the rebinding (Fig. 2*D*, *Left*). To test this prediction, we analyzed the differences in average pattern similarity between different triplet types (Fig. 2*D*, *Right*) using a repeated measures ANOVA (stimulus same vs. stimulus different, rebinding vs. no rebinding). We found no main effect of stimulus similarity [$F(1,27) = 1.451$, $P = 0.231$] or rebinding [$F(1,27) = 0.120$, $P = 0.73$]. Importantly, we found a significant interaction between the two factors [$F(1,27) = 9.091$, $P = 0.003$], indicating that rebinding reduced the effect on pattern similarity of whether the stimulus in trial $n - 2$ and n were the same or different. Follow-up t tests showed that there was a significant drop in similarity following rebinding when the same stimulus was shown on trials n and $n - 2$, i.e., AAA (M [mean] = 0.170, SD = 0.064) vs. ABA ($M = 0.161$, SD = 0.060), [$t(27) = 2.249$, $P = 0.033$]. The significant difference between "same" and "different" trials, i.e., AAA vs. BAA ($M = 0.155$, SD = 0.069) [$t(27) = 3.206$, $P = 0.0035$], representing classical coding, disappeared after rebinding (no significant difference for ABA and BBA ($M = 0.167$, SD = 0.065, [$t(27) = -1.090$, $P = 0.296$]). Interestingly, this was accompanied by a significant increase in similarity between "different" trials following a rebinding [i.e., BAA vs. BBA, $t(27) = -2.163$, $P = 0.040$]. These effects were unaffected by different ways of normalizing the data before analysis (*SI Appendix*, Fig. S2).

To confirm that this observed effect depended on trial history, we used a permutation test (Fig. 2*F*). Trials within a session were randomly reordered, and an identical analysis performed, to give a null distribution of t statistics. The main effect of "same" vs. "different" observed in the unshuffled data were not significantly higher than the ones observed in shuffled data, indicating that the classical coding was unaffected by trial order, i.e., was stable over time. However, the interaction in the unshuffled data were significantly larger ($P < 0.05$) than in the shuffled trials (monkey 1: $P < 0.1$, monkey 2: $P < 0.05$), confirming that the plastic code induced by rebinding was sensitive to trial order.

After establishing that the effect depends on trial history, we tested whether the signature of synaptic plasticity was present in the frontal eye fields (FEF) and the lateral intraparietal cortex (LIP) (Fig. 2*E*). We found no significant differences in pattern similarity between the different triplet types in either the FEF or the LIP ($P > 0.1$).

Changes in Code Are Not Due to Spillover from Previous Trial. The previous analysis demonstrated that the representation of a color—in terms of pattern similarity to a very recent trial—is disrupted when the trial immediately before it involves binding a different color to the same location. One explanation for this might be that activity related to the previous trial spills over into the current trial. Specifically, the cross-over interaction seen in the triplet analysis could arise if part of activation from trial $n - 1$ is carried over into trial n. If trial $n - 1$ contains the same stimulus as trial n, there is no contamination of the response in trial n. However, if trial $n - 1$ contains a stimulus that is different from trial n, any activation representing this stimulus carried into trial n will decrease similarity with trial $n - 2$ if the stimuli in trial n and $n - 2$ are the same and, conversely, increase similarity if the stimuli are different. Next, we address this possibility.

To test whether the observed effect could be caused by a carryover of signals from the previous to the current trial, we tested whether simply having a different feature on trial $n - 1$ (irrespective of whether this constitutes a rebinding of features) is sufficient to generate a drop in similarity. We repeated the triplet pattern similarity analysis while varying the distance between the first ($n - i$) and last trial (n) in the triplet. At distance $i = 2$, this matches the original triplet analysis [Fig. 3*A*; interaction between stimulus similarity and rebinding: $F(1,27) = 7.858$, $P = 0.006$]. At longer distances between $i = 3$ and $i = 100$ trials, there was a general drop in average similarity in all triplet types due to the intervening rebindings. However, we found that the presence of rebinding in specifically trial $n - 1$ was not sufficient to produce a drop in similarity between trials $n - i$ and n [Fig. 3*D*; $F(1,27) = 0.019$, $P = 0.892$]. This demonstrates that the rebinding effect cannot simply be due to activity from trial $n - 1$ spilling over into trial n. Rather, it is causally dependent on $n - 1$ being different from $n - 2$. As expected, the main effect of stimulus similarity was significant [$F(1,27) = 19.259$, $P < 0.001$] for these longer distances, confirming the presence of stable, classical stimulus identity coding over long periods of time.

Analysis of the time course in pattern similarity over the course of a trial showed that the stable fixed selectivity coding of color emerged ~200 ms after the presentation of the stimuli (Fig. 3*B*). This color encoding was observed more clearly for pairs of trials separated by longer distances (averaged across distances $n - 3$ to $n - 100$; Fig. 3*E*, $P < 0.05$ cluster-corrected), than for trial $n - 2$. The longer-distance analysis includes more trials and is therefore less noisy. Importantly, the signature of rebinding was only present in the average signal for trial distance of 2 (Fig. 3*C*), between 200 and 600 ms. However, none of the individual time point differences remained significant following cluster-based permutation correction for multiple comparisons. Interestingly, the opposite pattern emerged in the longer-distance analysis of rebinding (Fig. 3*F*). Here, repeated presentation of the same color in trials n and $n - 1$ generated a small, short-lived drop in pattern similarity compared to when a different trial was presented in trial $n - 1$. This pattern could potentially correspond to neural signature of adaptation to the previous trial. The results suggest that the rebinding effect cannot be solely due to carryover of activity from the previous trial.

Qualitatively, the pattern similarity difference between same- and different-stimulus trials appeared to reduce over about seven trials (*SI Appendix*, Fig. S8*B*), as per the predictions of a plastic model (*SI Appendix*, Fig. S8*A*). However, the difference remained strong when averaged over much longer periods of trials, indicating fixed selectivity that did not match the plastic model.

Representational Similarity Depends on Trial History. The preceding analyses indicate the presence of both plastic codes at short timescales and stable codes over longer periods. We can incorporate this into the model simulations, which yields new predictions: Short-term changes in selectivity should generate biases in the classical stable coding present across a session. In

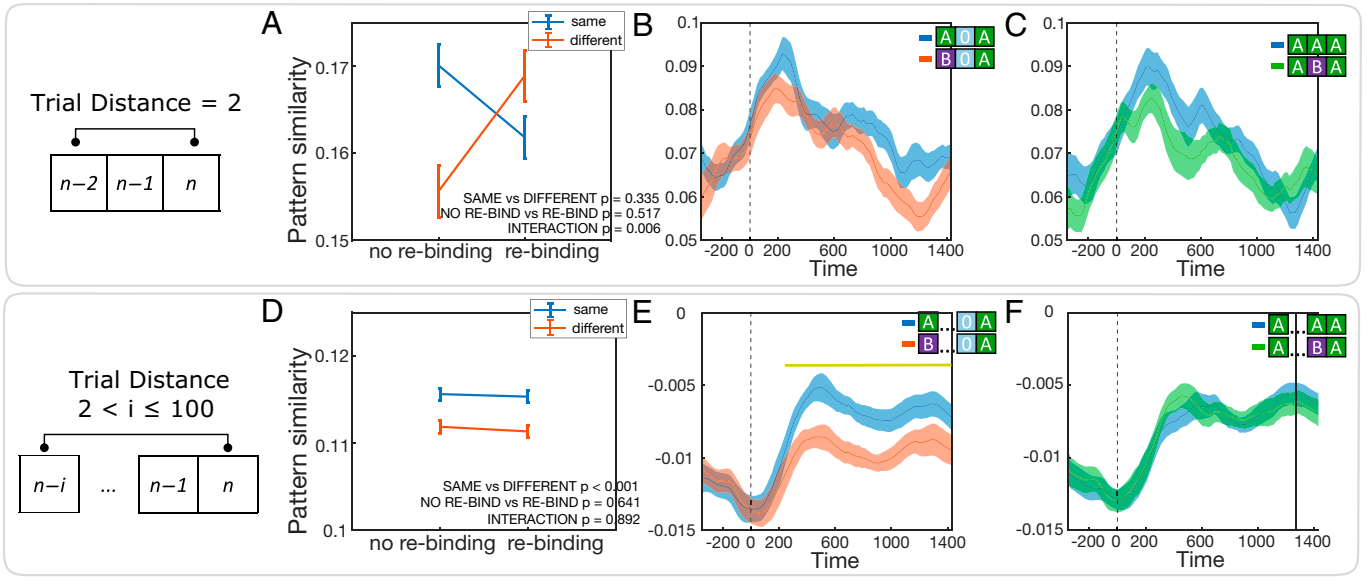

Fig. 3. Pattern similarity effect for different trial distances and over time–trial distance of 2 (*A–C*) and trial distances between 3 and 100 (*D* and *E*). (*A*) Triplet analysis showed that pattern similarity between trials representing the same stimulus dropped following a rebinding of association in trial $n - 1$. Importantly, this drop in similarity was not present when the first and last trial of the triplet were further apart in the session. (*B*) There was no stable coding of color over trials n and $n - 2$, whereas (*C*) the intervening trial had an impact on coding during the 200- to 600-ms poststimulus window, though individual timepoints were not significant. (*D*) In contrast to the distance of 2, rebinding no longer abolished the identity coding (i.e., main effect of same vs. different) for longer intertrial distances. (*E*) The difference in similarity of the neural pattern elicited by colors presented at the same location, i.e., stimulus identity coding, was visible from ~200 ms after the presentation of the stimuli in the signal from longer trial distances (blue versus red). (*F*) In contrast, the drop in similarity of neural response to the same stimulus after a rebinding of its color–location association was not present for longer trial distances (blue versus green). Significantly different time points are marked with a yellow bar (*$P < 0.05$; permutation-based correction for multiple comparisons).

other words, recent history interacts with background fixed selectivity in a consistent way across the session. To test these predictions, we compared similarity patterns in empirical data against simulations of five different modes of operation (see *Comparing Different Modes of Model Operation*): pure plasticity, pure fixed selectivity, mixed plastic and fixed selectivity, fixed selectivity with carryover from previous trial, and fixed selectivity with adaptation (Fig. 4). We simulated representation similarity matrices (RSM) for each. To obtain a value for each cell in the RSM, similarity was calculated for each pair of trials, and the pair was grouped based on the stimuli shown on those trials and the two preceding trials, i.e., what combination of triplets they came from. For example, the pattern similarity between a trial B from a triplet AAB and a trial A from triplet AAA occupies the very left bottom cell of the representational similarity matrix (RSM). This value corresponds to the average similarity of all pairs of trials of this type. The five different modes of operation resulted in specific patterns of similarity that could identify the contributions of different processes. For example, fixed selectivity yields identical patterns only for identical stimuli (i.e., last trial of each triplet); whereas a purely plastic-coding model yields low similarity when one of the trials requires rebinding, irrespective of whether the stimuli are the same or not.

To establish which of these signatures was present in the empirical data, a linear regression model was fitted to the data using the patterns generated by the five modes of operation as predictors, one at a time. The comparison of the individual models [fixed mode: $b = 0.32$, $t(56) = 6.847$, $P < 0.001$; plastic mode: $b = 0.16$, $t(56) = 2.010$, $P = 0.049$; fixed + plastic mode: $b = 0.73$, $t(56) = 7.353$, $P < 0.001$; fixed + adaptation mode: $b = 0.42$, $t(56) = 6.383$, $P < 0.001$; fixed + carryover mode: $b = 0.45$, $t(56) = 7.072$, $P < 0.001$] revealed that the plastic + fixed mode simulation was the best-fitting model for the empirical data ($BIC_{fixed+plastic} = -44.43$; the second-best-fitting model $BIC_{fixed+adaptive} = -42.28$, likelihood ratio = 8.58). Changing the amount of carryover or adaptation had little effect on the pattern similarity matrix (*SI Appendix*, Figs. S9 and S10). Introducing small amounts of fixed selectivity still allowed flexible coding, preserving the pattern similarity interaction between rebinding and stimulus identity (*SI Appendix*, Fig. S11).

One analysis examined stimulus encoding at only one location at a time. According to the model, changes at other locations might increase code shifts due to rebinding, but this was not observed in the data (*SI Appendix*, Fig. S6).

Discussion

The ability of neurons to adaptively represent new associations is considered essential for flexible behavior. Rapid synaptic plasticity has been proposed as one of the potential mechanisms allowing neurons to code novel, task-relevant combinations of features over short timescales. This could result in changes in neural selectivity at the trial-to-trial timescale. Here, we analyzed existing data looking for changes in coding of neural populations in the PFC. We showed that the neural population code for a particular stimulus changes between trials if the feature association within the stimulus is broken in an intervening trial.

We found evidence for a significant drop in pattern similarity between neural responses to the same stimulus if separated by a different stimulus that violates the features' binding. We hypothesize that this shift in neural response is explainable in terms of rapid changes in synaptic efficacy (24). If rapid Hebbian plasticity is acting on the connections between sensory neurons coding for individual features of the item (i.e., location and color) and PFC neurons that bind features together, then violation of original association alters the connection strengths in such a way as to weaken previously established selectivity within the pattern of synaptic weights. Due to this erosion of the original pattern of coding, encoding the original association

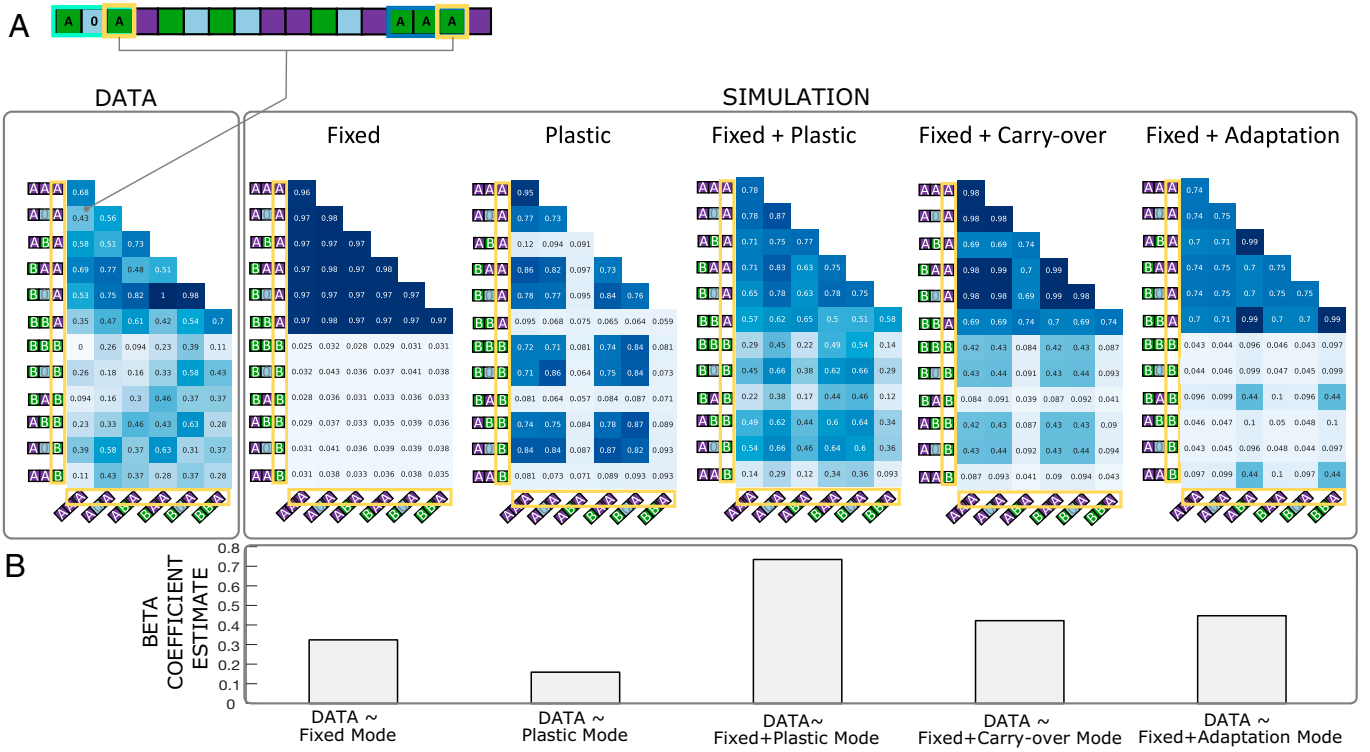

Fig. 4. Scaled representational similarity matrices and beta coefficient estimates from the individual linear regression analyses. (*A*) Representational similarity matrices for data and different modes of operation. Each square within the matrix corresponds to the average correlation between the last trials (framed in yellow) of two triplet types. For example, the second square from the top represents a correlation between the last trial (trial A) from triplet ABA and the last trial (trial A) from the triplet AAA. Different modes of operation of the model resulted in specific patterns representing different types of processes contributing to neural coding. For example, the distinct half-split pattern in the fixed mode corresponds to traditional fixed selectivity, in which trials that contained the same stimuli have high correlation in neural response (dark blue) whereas trials that contained different stimuli were uncorrelated (light blue). Most importantly, there is no variation in this fixed selectivity as a function of trial history (i.e., what trials preceded the trials in question) in this mode of operation. (*B*) Regression coefficient estimates for each of the modes of operation.

again will be influenced by the changes in connections caused by the rebinding. As a result, the new pattern of response may be more dissimilar to the original neural response (Fig. 2*D*, *Right*). This pattern of results was only present in the PFC neurons and not in LIP or FEF.

Short-term, Hebbian plasticity-based accounts of neural coding posit that the neural response to a particular stimulus depends on the history of events the neural population coded previously. This does not mean that neurons are selective for a particular history of events. Instead, a neuron's response to an object depends on the interplay between prior synaptic state and the particular combination of inputs. The need to reassociate features might mean that a neuron becomes either more or less likely to respond to a given feature, depending on how well that neuron and other neurons can accommodate the new association. Our novel triplet pattern similarity analysis was aimed at discovering these changes in coding as a function of changes in feature associations. Previous studies have demonstrated drifting neural codes (29, 30) but these do not control for time itself, where shifts in the recorded neurons may accumulate over time. By looking at short-term changes in coding over the course of three trials, we could uncover changes in selectivity while controlling for these confounds. In our analysis, changes in tuning are not simply due to drift; they are stimulus-driven (see Fig. 2). The modeling also demonstrates how other neurons can still read out information, even when tuning is flexible. The plasticity is bidirectional, so flexible PFC neurons become the coordinating node of a new attractor state with other brain areas. The symmetric Hebbian rule acts so that both the input and output synapses of these flexible neurons change together. This ensures a stable readout, in the form of pattern completion or reactivation.

Some features of the neural data are not as prominent in the model, such as the crossover interaction. In particular, not only did rebinding decrease the similarity between two trials coding for the same stimulus, it also increased similarity between two trials coding a different stimulus. In this experimental setup, this means that the neural response to a current, red stimulus presented in the left upper location is more similar to a trial in which a green stimulus was in the same location, if these two trials were separated by a trial with a green stimulus in that location (Fig. 2*D*, *Right*). The simulation does predict this increase in similarity for different stimuli (i.e., positive slope) following rebinding. This is because repeated presentation of the green stimulus in the upper-right location generates a stronger synaptic trace between the feature units and the conjunction unit coding for this binding. Due to this increase in connection strength, the activation from feature units coding a new association, i.e., the red stimulus at that location, is more likely to be guided down these connections and, as a result, encoded by the same conjunction unit. This in turn produces higher similarity in the neural responses between the two trials containing a different stimulus. One potential outcome of this could be that this conjunction unit acquires a strong connection to both green and red colors, which could result in more confusion during recall. Indeed, on average, the monkeys tended to make more errors after rebinding (*SI Appendix*, Fig. S4), although this effect could also be explained in terms of stimulus-level carryover.

An alternative explanation for this cross-over interaction seen in the triplet analysis is that part of activation from trial $n-1$

is carried over into trial *n*. To address this possibility, we demonstrated that the change in similarity was only present within triplets and was not present between pairs of trials with distances between 2 and 100 trials. For trials that were further apart in time, we observed a signature of fixed selectivity—a clear separation in pattern similarity between pairs of trials containing two of the same stimuli versus two different stimuli. This finding is analogous to standard decoding analysis (24).

The presence of fixed selectivity at longer distances that is abolished when triplets of successive trials are analyzed could be partly explained if plastic changes to the neural response happen on top of more sustained neural selectivity. Rather than history-driven changes in selectivity completely abolishing neural selectivity over time, as would arise with fully nonselective weights, it is likely that changes in synaptic efficacy only partially alter the population selectivity. To address this possibility, we simulated modes of operation that rely only on fixed selectivity or use a combination of fixed and plastic weights, where synaptic efficacy is the sum of fixed and Hebbian components. We also tested for the contribution of two other processes, neural adaptation and carryover. We were able to show that a combined fixed and plastic selectivity mode of model operation was the best at fitting the pattern of results in the empirical data. This suggests that the population response is a function of multiple neural mechanisms including ones that are in line with synaptic plasticity as well as fixed selectivity. Such mixed architectures have been proposed to account for multiple timescales of retention in WM (31). In FEF, there was a suggestion of an interaction between the identity of the stimulus and rebinding, but this was not significant, potentially due to fewer data in FEF (*SI Appendix*, Fig. S3).

In summary, we devised a model-based analysis method, which we applied to single-neuron data to provide evidence for short-term, task-history-dependent changes in neural coding. These changes were in line with the predictions of rapid Hebbian synaptic plasticity and were only present in the PFC. The findings may help to explain why in some situations, stable stimulus selectivity in PFC is generally found to be weak, despite neurons' being consistently active during memory tasks. Flexible tuning means that coding need not be stable, even though these neurons play a vital role in binding of features.

Methods

Experimental Task and Data. The experimental data used in this study were originally collected and analyzed in a study by Buschman et al. (28). Two adult rhesus monkeys (*Macaca mulatta*) performed a change localization task (Fig. 1*B*) while simultaneous recordings were made from single neurons in PFC (lateral PFC, 584 neurons; FEF, 325 neurons) and parietal cortex (LIP, 284 neurons; see the supplemental materials for ref. 28 for further details on recordings and spike extraction). Each monkey completed 14 recording sessions.

In the beginning of each trial, monkeys fixated on a central fixation point for 500 ms. If fixation was broken at any point during the trial before response cue, the trial was marked as aborted. The fixation period was followed by 800-ms-long presentation of sample display containing colored squares presented at six possible locations (±75 angular degrees from the horizontal meridian and between 4° and 6° of visual angle from fixation). Two possible colors for each of the six locations as well as the locations themselves were randomly chosen at the beginning of each recording session. The number of stimuli varied from trial to trial between two and five items. This meant that on a given trial, a given location could contain either of two colors or be empty. A memory delay of varying length (800 to 1,000 ms) was followed by a test array that was identical to the sample display except the color of one of the squares changed. The monkeys were trained to make an eye movement toward the square that changed its color between sample and test.

Data Preprocessing. Data from each session were first arranged into spike–time sequences (neurons × trials × time points) and cropped into three trial epochs: prestimulus interval (500 ms), stimulus presentation (~800 ms), and delay period (~800 to 1,000 ms). For each trial, each neuron's average firing rate (FR) was calculated as the sum of spikes within the stimulus presentation and delay periods divided by time (i.e., length of the epoch in milliseconds). For temporal analyses, FRs were calculated for each trial and neuron over a temporal window of 300 ms in steps of 20 ms.

Normalization. FRs from each session were normalized in the following way. First, each neuron's FR was z-scored across trials, and then for each trial the FRs were z-scored across neurons. The first step centered the FR of each neuron at zero across trials, providing a measure of the neuron's activation on each trial relative to all the other trials. The second step then transformed a neuron's FR on a given trial to represent its activity level with respect to all the other neurons, to give a FR pattern centered on zero. The purpose of the second transformation was to center the activation more locally compared to the more global, across-session, normalization.

We also confirmed that these normalization approaches did not make a difference to the primary finding, with qualitatively similar results when only step 1 or step 2 were performed alone, or when both were applied in reverse order (*SI Appendix*, Fig. S2).

Grouping of trials. Throughout the analysis, we examined just one of the spatial locations at a time. On a given trial, the location contained one of the two colors or was empty. This gives three possible stimuli, which we denote A, B, or O. For each session, conditions were determined using these labels, and the analysis was repeated using the labels for each of the six locations. The results for each location were then pooled and averaged.

Model and Simulated Data. All simulations were performed using an existing computational model of WM [*SI Appendix*, Fig. S1; Manohar et al. (24)]. The simulated data and scripts that support the findings in this study are openly available at https://osf.io/fmxs4/?view_only=282caf168da643a882d689fcff15c26f.

We focus on encoding a single item in the display into WM, which in the task corresponds to the colors shown at just one of the spatial locations. The model contains 12 feature-selective units, arranged into three groups of four and represented by activation (f). Each of these units is connected to a group of four freely conjunctive units (c). Each unit is self-excitatory and activity is constrained between 0 and 1. Interaction between units within each group is guided by blanket lateral inhibition. The full activity update equations are the following:

$$c \leftarrow \sigma(\beta + (\alpha_1 W^{cc} + \alpha_2 I)(c - \beta) + \alpha_3 W^{cf}(f - \beta) + \varepsilon \cdot N) \quad [1]$$

$$f \leftarrow \sigma(\beta + (\alpha_4 W^{ff} + \alpha_5 I)(f - \beta) + \alpha_6 W^{fc}(c - \beta) + i), \quad [2]$$

where c and f are the units' FRs and β is the units' baseline activity; α_1, α_4 are free parameters for mutual lateral inhibition between neurons; α_2, α_5 are the self-excitation parameters; I are identity matrices; α_3, α_6 denote synaptic gain; $\varepsilon \cdot N$ corresponds to Gaussian white noise (SD 1); and σ is a function constraining values between 0 and 1. W_{cf} and W_{fc} are conjunctive-feature synapses [initiated as pseudorandom values drawn from the standard uniform distribution on the open interval (0,1)] and W_{ff} and W_{cc} are fixed interfeature and interconjunction synapses respectively (1 for synapses within dimension, 0 otherwise). The synapses between the feature and conjunctive neurons are continuously updated by a Hebbian covariance rule. This allows patterns of inputs to be encoded into conjunctive units:

$$\Delta = (c - \beta) \cdot (f^T - \beta) \quad [3]$$

$$W \leftarrow \sigma(W + \gamma\Delta), \quad [4]$$

where Δ is the change in FR and γ is the learning rate.

One model parameter was slightly adjusted to allow the model to perform under different modes of operation (see below). Specifically, the self-excitation (α_2) in conjunctive unit activity was changed from 1 to 0.98. The meanings of the 12 feature-selective units were mapped to the behavioral task in the following way. The first dimension represented color, the second dimension represented location, and the third dimension represented context. The plastic attractor network therefore functions to associate a given color with a given location, while accounting for the possibility that other items on the display induce a varying context.

To simulate the behavioral task, each trial started with a prestimulus period lasting 200 time steps (ts) with all features under maximal inhibition (input i = −1). During encoding (120 ts), input (i = +1) into one randomly chosen color was provided together with its assigned location (two colors were associated with one location as in the empirical experiment) and one randomly chosen context out of four possible context neurons. The random context reflects the fact that colors are also presented at the nonsimulated locations in the empirical experiment. All other features were suppressed (i = −1). After a memory delay period (300 ts; i = 0) the encoded location was activated (120 ts; i = +1, i = −1 for all other feature units) followed by a period of no input (220 ts; i = 0). The model was exposed to these trial sequences in blocks of 2,000 trials repeated over 20 simulations. The average activity during the delay period was used as the equivalent of FR calculated for the neural data. Simulated activity was normalized in the same way as described above.

Comparing different modes of model operation. To simulate the different possible mechanisms contributing to changes in coding, the model was run using three different weight settings: 1) plastic weights only, 2) fixed weights only, and 3) plastic and fixed weights combined. The first mode (i.e., plastic mode) corresponded to the basic operation of the model as described above and in Manohar et al. (24). In the second mode that corresponded to pure fixed selectivity, we disabled the plastic changes to the weights, which were instead fixed at random uniform values between 0 and 0.5. In the third, combined mode, a stable set of fixed weights (same weights as for the fixed weights only mode) was added to the ongoing plastic changes to the synaptic weights as follows: in Eq. **1** W^{cf} was replaced with $(W^{cf} + W^{fixed})$ and Eq. **2** W^{fc} was replaced with $(W^{fc} + W^{fixed\,T})$. Two additional modes of operation were generated by modifying the results of the fixed weights only simulation. To simulate possible carryover of activity from trial $n - 1$ to trial n, we calculated the previous trial's average activity for each C-unit during the probe period and added 20% of this activity pattern to the activity in the current trial. Conversely, to simulate adaptation, we subtracted 20% of the previous trial's activity from the current trial.

Twenty sets of random fixed weights were used, and each was simulated 20 times.

Pattern Similarity. To evaluate changes in neuronal coding between trials, pattern similarity was calculated using normalized FR from pairs of trials of interest (Fig. 2*C*). The similarity between the vector of FRs for trial n and trial $n - 2$ was computed using cosine similarity:

$$similarity(y_n, y_{n-2}) = \frac{y_n \cdot y_{n-2}}{\|y_n\| \times \|y_{n-2}\|},$$

where y_n and y_{n-2} are vectors of FRs from trials n and $n - 2$ and $\| \|$ corresponds to the magnitude of the vector. Cosine similarity represents the cosine of the angle between two vectors in multidimensional space.

Triplet pattern similarity. To test the hypothesis that rebinding of features results in a change in neural coding, triplets of consecutive trials were assigned to six different triplet types based on feature similarity between trials n and $n - 2$ (i.e., same versus different color) and the presence of binding rearrangement between trial n and trial $n - 1$ (rebinding versus no rebinding versus absent if no stimulus was presented at the location): same no rebinding, same rebinding, same absent, different no rebinding, different rebinding, and different absent (Fig. 2*A*). Any neurons with missing values within the triplet of interest were removed and pattern similarity was calculated as described above between trial n and trial $n - 2$.

Triplet permutation analysis. We performed permutation analysis to test whether any significant differences in pattern similarity obtained through the triplet pattern similarity analysis described above were dependent on the temporally consecutive arrangement of the trials within the triplets. To do this, we randomly shuffled the trials together with their labels (i.e., trial label assignments were maintained) 1,000 times. Triplet selection followed by pattern similarity calculation was repeated for each of these iterations. A t value for the main effect of color, main effect of rebinding, and their interaction was then calculated for each of the iterations generating a null distribution of the t values for each of these effects. The interaction t value was calculated in the following way: the difference between the same, no rebinding and different, no rebinding conditions was compared against the difference between same, rebinding and different, rebinding conditions. The t values from the unshuffled data were then compared against this null distribution to determine statistical significance of these effects. If real t value was larger than 95% of the t values from the null distribution, the effect was considered significant at $P < 0.05$.

Pattern similarity between more distant trials. To check whether any drop in similarity between trials $n - 2$ and n was simply a result of activation related to the stimulus from the previous trial being carried into the current trial, we varied the distance between the first and last trial of each triplet between 3 and 100 trials and calculated pattern similarity as described above.

Pattern similarity over time. To examine the temporal evolution of the differences between trial triplets over the course of a trial, we calculated pattern similarity over time. In this case, pattern similarity for each triplet type was calculated using FRs from consecutive time windows (see *Data Preprocessing*). The resulting pattern similarity was then further smoothed (step size 5 resulting in ∼400-ms smoothing) and differences between triplets of interest at each time point were calculated using t tests. The resulting P values were corrected for multiple comparisons using a permutation test.

Effect of Trial History on Stable Coding, Using Representational Similarity. Representational similarity analysis was performed to examine how temporary local changes in coding affect more stable color identity coding observed on the level of a session. In this analysis, triplets of interest for each of the possible feature values (i.e., color 1- same, no rebinding; color 2- same, no rebinding, etc.) were identified as described in *Triplet Pattern Similarity*. Any neurons with missing values within any of the trials n of the identified triplets were removed. FRs, normalized as described in *Normalization*, from the last trial in each triplet were then used to calculate pattern similarity and grouped based on the type of trial triplets they came from. This produced 12 classes of trial history, depending on which stimulus color was shown on trial n, $n - 1$, and $n - 2$ of both trials in the pair (Fig. 4). The final RSM contained the average cosine similarity for all possible combinations of the triplets' last trials and as a result was symmetrical along its diagonal. To remove these repetitions, one side of the RSM along its diagonal was ignored. Correlations corresponding to the same conditions with swapped feature values (e.g., same no rebinding irrespective of whether the triplet in question came from color 1 or color 2) were averaged. This produced a RSM with 57 elements, showing the effect of trial history on coding of color.

RSMs were calculated separately 20 times, each time with a different set of random weight matrices, and then averaged to generate an average RSM matrix for each of the modes of operation.

Regression analysis. We asked which of the model's modes of operation (plastic, fixed, fixed + plastic, fixed + adaptive, fixed + carryover) generated a pattern of data that was the most predictive of real data. The elements of the RSM matrix from empirical data were compared to the RSMs from simulations of the five different modes. Prior to each regression, each matrix was zero-centered and scaled to give a variance of 1, to allow comparison between the different modes. We used five separate linear regressions (MATLAB function fitlm) to express the data similarity pattern, a 57-element vector, as a function of the corresponding patterns seen in the individual simulations. We then compared the beta coefficient estimates from the individual linear models and used Bayesian information criterion to select the best fitting model for the empirical data.

Data, Materials, and Software Availability. All model code and results of simulations have been deposited in OSF (32) (https://osf.io/fmxs4/?view_only=282caf168da643a882d689fcff15c26f. Experimental data available upon reasonable request from the authors.

1. J. Duncan, An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820–829 (2001).
2. A. Baddeley, The concept of working memory: A view of its current state and probable future development. *Cognition* **10**, 17–23 (1981).
3. M. D'Esposito, B. R. Postle, The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).
4. E. K. Miller, The "working" of working memory. *Dialogues Clin. Neurosci.* **15**, 411–418 (2013).
5. K. Oberauer, "Design for a working memory" in *Psychology of Learning and Motivation*, B. Ross, Ed. (Academic Press, 2009), **vol. 51**, pp. 45–100.
6. F. Bouchacourt, T. J. Buschman, A flexible model of working memory. *Neuron* **103**, 147–160.e8 (2019).
7. S. L. Franconeri, G. A. Alvarez, P. Cavanagh, Flexible cognitive resources: Competitive content maps for attention and memory. *Trends Cogn. Sci.* **17**, 134–141 (2013).

8. K. K. Sreenivasan, C. E. Curtis, M. D'Esposito, Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18**, 82–89 (2014).
9. M. G. Stokes, 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
10. M. J. Wolff, J. Jochim, E. G. Akyürek, T. J. Buschman, M. G. Stokes, Drifting codes within a stable coding scheme for working memory. *PLoS Biol.* **18**, e3000625 (2020).
11. B. Scholl, D. Fitzpatrick, Cortical synaptic architecture supports flexible sensory computations. *Curr. Opin. Neurobiol.* **64**, 41–45 (2020).
12. S. V. David, B. Y. Hayden, J. A. Mazer, J. L. Gallant, Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* **59**, 509–521 (2008).
13. M. Siegel, T. J. Buschman, E. K. Miller, Cortical information flow during flexible sensorimotor decisions. *Science* **348**, 1352–1355 (2015).
14. A. Woolgar, S. Afshar, M. A. Williams, A. N. Rich, Flexible coding of task rules in frontoparietal cortex: An adaptive system for flexible cognitive control. *J. Cogn. Neurosci.* **27**, 1895–1911 (2015).
15. G. Mongillo, O. Barak, M. Tsodyks, Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
16. Y. Wang *et al.*, Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
17. R. S. Zucker, W. G. Regehr, Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
18. F. Fiebig, A. Lansner, A spiking working memory model based on Hebbian short-term potentiation. *J. Neurosci.* **37**, 83–96 (2017).
19. M. A. Erickson, L. A. Maramara, J. Lisman, A single brief burst induces GluR1-dependent associative short-term potentiation: A potential mechanism for short-term memory. *J. Cogn. Neurosci.* **22**, 2530–2540 (2010).
20. P. Park *et al.*, NMDA receptor-dependent long-term potentiation comprises a family of temporally overlapping forms of synaptic plasticity that are induced by different patterns of stimulation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130131 (2013).
21. A. Volianskis, M. S. Jensen, Transient and sustained types of long-term potentiation in the CA1 area of the rat hippocampus. *J. Physiol.* **550**, 459–492 (2003).
22. F. Fiebig, P. Herman, A. Lansner, An indexing theory for working memory based on fast Hebbian plasticity. BioRxiv [Preprint] (2018). https://www.biorxiv.org/content/10.1101/334821v7 (Accessed 24 Jun 2021).
23. Q.-S. Huang, H. Wei, A computational model of working memory based on spike-timing-dependent plasticity. *Front. Comput. Neurosci.* **15**, 630999 (2021).
24. S. G. Manohar, N. Zokaei, S. J. Fallon, T. P. Vogels, M. Husain, Neural mechanisms of attending to items in working memory. *Neurosci. Biobehav. Rev.* **101**, 1–12 (2019).
25. S. M. Polyn, K. A. Norman, M. J. Kahana, A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.* **116**, 129–156 (2009).
26. M. Rigotti *et al.*, The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
27. G. Swan, B. Wyble, The binding pool: A model of shared neural resources for distinct items in visual working memory. *Atten. Percept. Psychophys.* **76**, 2136–2157 (2014).
28. T. J. Buschman, M. Siegel, J. E. Roy, E. K. Miller, Neural substrates of cognitive capacity limitations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11252–11255 (2011).
29. S. E. Cavanagh, J. D. Wallis, S. W. Kennerley, L. T. Hunt, Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. *eLife* **5**, e18937 (2016).
30. M. E. Rule *et al.*, Stable task information from an unstable neural population. *eLife* **9**, e51121 (2020).
31. H. Lee, W. Choi, Y. Park, S.-B. Paik, Distinct role of flexible and stable encodings in sequential working memory. *Neural Netw.* **121**, 419–429 (2020).
32. A. Bocincova, T. J. Buschman, M. G. Stokes, S. G. Manohar, Neural signature of flexible coding in prefrontal cortex. OSF. https://osf.io/fmxs4/?view_only=282caf168da643a882d689fcff15c26f, doi:10.17605/OSF.IO/FMXS4. Deposited 16 May 2022.