nature neuroscience

Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information

Marino Pagan, Luke S Urban, Margot P Wohl & Nicole C Rust

Finding sought visual targets requires our brains to flexibly combine working memory information about what we are looking for with visual information about what we are looking at. To investigate the neural computations involved in finding visual targets, we recorded neural responses in inferotemporal cortex (IT) and perirhinal cortex (PRH) as macaque monkeys performed a task that required them to find targets in sequences of distractors. We found similar amounts of total task-specific information in both areas; however, information about whether a target was in view was more accessible using a linear read-out or, equivalently, was more untangled in PRH. Consistent with the flow of information from IT to PRH, we also found that task-relevant information arrived earlier in IT. PRH responses were well-described by a functional model in which computations in PRH untangle input from IT by combining neurons with asymmetric tuning correlations for target matches and distractors.

Searching for a specific object, such as your car keys, begins by activating and maintaining a representation of your target in working memory. Finding your target requires you to compare the visual content of a currently viewed scene with this working memory representation to determine whether your target is currently in view. Our ability to rapidly and robustly switch between different targets suggests that this process is highly flexible. How do our brains achieve this?

Theoretical proposals of how our brains might find objects and switch between targets differ in their details[1–4], but all propose that visual and target-specific working memory signals are first combined to produce a target-modulated visual representation, followed by a second stage in which the combined signals are reformatted to produce a signal that reports when a currently viewed scene contains a target (**Fig. 1**). However, the means by which these signals are combined and reformatted remains poorly understood. Working memory signals are thought to be maintained in higher order structures, such as prefrontal cortex (PFC), and these signals are thought to be fed back to earlier structures for combination with visual information[3,5,6] (but see ref. 4). The initial combination of visual and working memory signals is likely to occur in higher stages of the ventral visual pathway (for example, V4 and IT) via a process known as feature-based, or object-based, attention, as evidenced by V4 and IT neurons whose responses are modulated by both the identity of the visual stimulus and the identity of a sought target[7–14]. Although many models incorporate the simplifying assumption that the initial combination is implemented similarly by all neurons (for example, a multiplicative enhancement aligned with a neuron's preferred visual stimulus), experimental evidence suggests that these initial mechanisms are in fact quite heterogeneous[7,8,11,13,15]. These little-understood rules of combination likely determine the computations that the brain subsequently uses to determine whether a target is present in a currently viewed scene.

To explore how visual and working memory signals are combined, we trained macaque monkeys to perform a well-controlled, yet simplified, version of target search in the form of a delayed match-to-sample task that required them to sequentially view images and respond when a target image appeared. Our experimental design required them to treat the same images as targets and distractors in different blocks of trials. As monkeys performed this task, we recorded responses in IT, the highest stage of the ventral visual pathway. Our results suggest that visual and working memory signals are combined in a heterogeneous manner and the result is a nonlinearly separable, or tangled[16], IT representation of whether a target is currently in view. To explore the computations by which this type of representation is transformed into a report of whether a target is present, we also recorded signals in PRH, which receives its primary input from IT[17] and has been found in lesioning studies to be important for visual target search tasks[18] (but see ref. 19). Our results indicate that information about whether a target is currently in view is more untangled[16] or more linearly separable in PRH and that the PRH population representation differs on correct as compared with error trials. Models fit to our data revealed that the responses of neurons in PRH are well-described by an untangling process that works by combining signals from IT neurons that have asymmetric tuning correlations for target matches and distractors (for example, have similar tuning for target matches and anti-correlated tuning for distractors).

RESULTS

IT and PRH responses are heterogeneous

We recorded neural responses in IT and PRH as monkeys performed a delayed match-to-sample, sequential object search task that required them to treat the same images as targets and distractors in different blocks of trials (**Fig. 2a**). Behavioral performance was high overall (monkey 1, 94% correct; monkey 2, 92% correct; **Supplementary Fig. 1a**). Performance remained high on trials that included the same distractor presented repeatedly before the target match (monkey 1, 89% correct; monkey 2, 86% correct), confirming that the monkeys

Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to N.C.R. (nrust@sas.upenn.edu).

Received 9 January; accepted 8 May; published online 23 June 2013; doi:10.1038/nn.3433

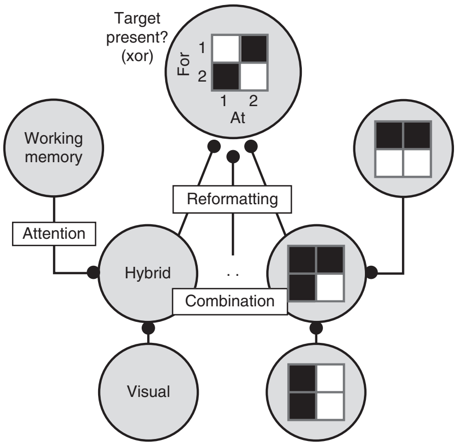


Figure 1 Theoretical proposals of the neural mechanisms involved in finding visual targets. Theoretical models propose that visual signals and working memory signals are nonlinearly combined in a distributed fashion across a population of neurons, followed by a reformatting process to produce neurons that explicitly report whether a target is present in a currently viewed scene. The delayed match-to-sample task is logically equivalent to the inverse of an 'exclusive or' (xor) operation in that the solution requires a signal that identifies target matches as the conjunction of looking at and for the same object. Shown (top) is a theoretical example of such a 'target present?' neuron, which fires when ('at', 'for') is (1,1) or (2,2), but not (1,2) or (2,1). Producing such a signal requires at least two stages of processing in a feedforward network[40]. As a simple example, a 'target present?' neuron could be constructed by first combining visual and working memory inputs in a multiplicative fashion to produce hybrid detectors that fire when individual objects are present as targets, followed by pooling. Note that this is not a unique solution.

were generally looking for specific images as opposed to detecting the repeated presentation of any image (consistent with ref. 15). Altogether, we presented four images in all possible combinations as a visual stimulus (looking at) and as a target (looking for), resulting in a four-by-four response matrix (**Fig. 2b**). As monkeys performed this task, we recorded neural responses in IT and PRH. To examine response properties, we counted spikes after the onset of each test (that is, non-cue) stimulus in a window that accounted for neural latency, but also preceded the monkeys' reaction times (80–270 ms; Online Methods and **Supplementary Fig. 1b**), unless otherwise stated. We then screened for neurons that were significantly modulated across the 16 conditions, as assessed by a one-way ANOVA (Online Methods). Unless otherwise stated, our analyses were based on the data from correct trials.

The three components of this task produce distinct structure in these response matrices: visual selectivity translates to vertical structure (**Fig. 2b**), working memory selectivity for the current target translates to horizontal structure (**Fig. 2b**), and, because matches fall along the diagonal of this matrix and distractors fall off the diagonal, differential responses to target matches and distractors translate to diagonal structure (four-object target detector, single-object target detector and suppressed four-object target detector; **Fig. 2b**). We found the four-object target detectors particularly compelling, as their matrix structure reflects the solution to the monkeys' task (that is, these neurons fired differentially when an image was viewed as a target versus as a distractor, and they did so for all four images included in the experiment). We also note that these examples of relatively pure selectivity existed in IT and PRH populations that were largely heterogeneous mixtures of different types of information. Note, for example, the distractor detector, which fired when image 2 was the stimulus and image 3 was the target as well as the mixture neuron (**Fig. 2b**).

PRH contains more untangled target match information

How do the heterogeneous responses of IT and PRH neurons relate to a determination of whether a currently viewed image matches the sought target (that is, the solution to the monkey's task)? To assess this relationship, we began by probing the amount of untangled target match information in the IT and PRH populations with

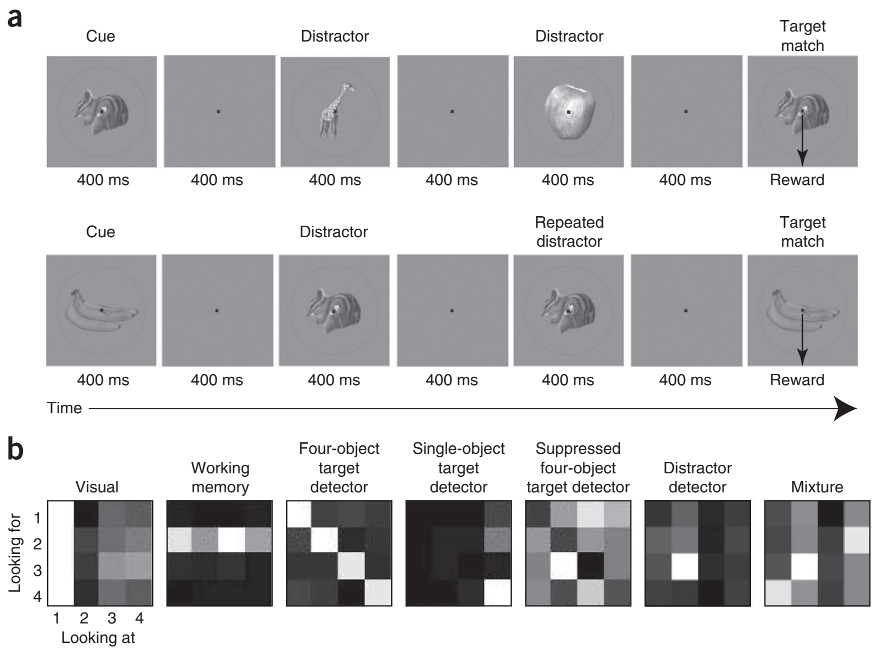


Figure 2 The delayed match-to-sample task and example neural responses. (**a**) We trained monkeys to perform a delayed match-to-sample task that required them to treat the same four images (shown here) as target matches and as distractors in different blocks of trials. Monkeys initiated a trial by fixating a small dot. After a delay, an image indicating the target was presented, followed by a random number (0–3, uniformly distributed) of distractors, and then the target match. Monkeys were required to maintain fixation throughout the distractors and make a downward saccade when the target appeared to receive a reward. Approximately 25% of trials included the repeated presentation of the same distractor with zero or one intervening distractors of a different identity. (**b**) Each of four images were presented in all possible combinations as a visual stimulus (looking at), and as a target (looking for), resulting in a four-by-four response matrix. Shown are the response matrices for example neurons with different types of structure (labeled). All matrices depict a neuron's response with pixel intensity proportional to firing rate, normalized to range from black (the minimum) to white (the maximum) response. We recorded these example neurons in the following brain areas (left to right): PRH, PRH, PRH, IT, PRH, IT and IT. Single-neuron linearly separable information (I_L; **Fig. 4c**) values (left to right) were 0.01, 0.02, 3.33, 0.39, 0.44, 0.01 and 0.06.

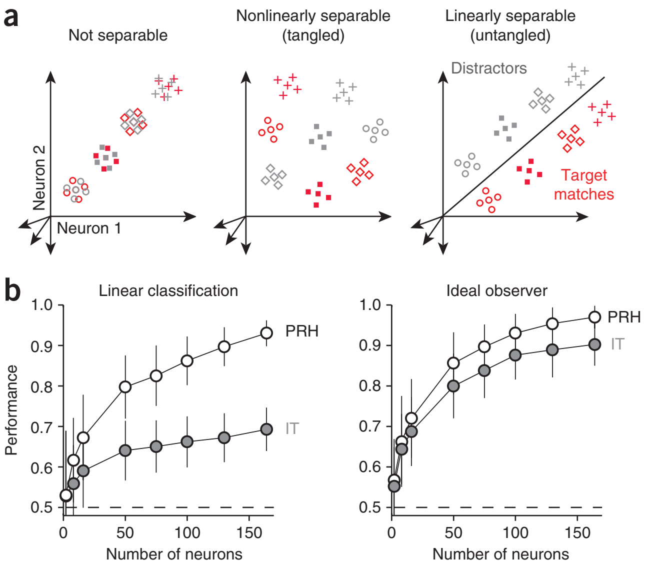


Figure 3 Population performance. (**a**) Each point depicts a hypothetical population response, consisting of a vector of the spike count responses to a single condition on a single trial. The four different shapes depict the hypothetical responses to the four different images and the two colors (red, gray) depict the hypothetical responses to target matches and distractors, respectively. For simplicity, only 4 of the 12 possible distractors are depicted. Clouds of points depict the predicted dispersion across repeated presentations of the same condition as a result of trial-by-trial variability. The target-switching task (**Fig. 2**) required discriminating the same objects presented as target matches and as distractors. (**b**) Performance of the IT (gray) and PRH (white) populations, plotted as a function of the number of neurons included in each population, via cross-validated analyses designed to probe linear separability (left) and total separability (linear and/or nonlinear, right). The dashed line indicates chance performance. We measured linear separability with a cross-validated analysis that determined how well a linear decision boundary could separate target matches and distractors. We measured total separability with a cross-validated, ideal observer analysis. Error bars correspond to the standard error that can be attributed to the random assignment of training and testing trials in the cross-validation procedure and, for populations smaller than the full data set, to the random selection of neurons.

a linear read-out (**Fig. 3a**). More specifically, we determined how well a linear decision boundary could separate target matches from distractors via a cross-validated analysis that involved using a subset of the data to find the linear decision boundary via a machine learning procedure (support vector machine) and then tested the boundary with separately measured trials (equation (1), Online Methods). Cross-validated population performance was significantly higher in PRH than in IT ($P < 0.005$; **Fig. 3b**) and this result was confirmed in each monkey individually (**Supplementary Fig. 2a**). Higher PRH performance could not be explained by the repeated presentation of the match after it had previously been presented in the trial as the cue or by changes in reward expectation as a function of the number of distractors encountered thus far in a trial or other position effects (**Supplementary Fig. 2a**). Finally, although these analyses assume trial-by-trial independence between neurons, correlated variability has been shown to affect linear read-out population performance for some tasks[20,21]. For our data, we tested the independence assumption by analyzing smaller subpopulations of simultaneously recorded neurons, and found similar results when the noise correlations were kept intact and when they were scrambled (**Supplementary Fig. 2b**).

We were also interested in determining whether our recorded responses were consistent with a putative role in the circuitry that transforms sensory information into a behavioral response. Consistent with this hypothesis, PRH linear classification performance peaked well before the monkeys' behavioral reaction times, which were longer than 270 ms on these trials (**Fig. 4a** and **Supplementary Fig. 2c**). We also found that linear classification performance on error trials, as compared with correct trials, trended toward lower values in IT ($P = 0.11$) and was significantly lower in PRH ($P = 0.009$; **Fig. 4b**). Poorer error trial performance could not be attributed simply to a difference in firing rate (grand mean firing rates: IT correct = 7.6 Hz, error = 7.2 Hz, $P = 0.26$; PRH correct = 5.6 Hz; error = 5.5 Hz, $P = 0.45$).

What response properties can account for higher performance in the PRH as compared with the IT population? We computed a single-neuron measure of linearly separable target match information (I_L) as

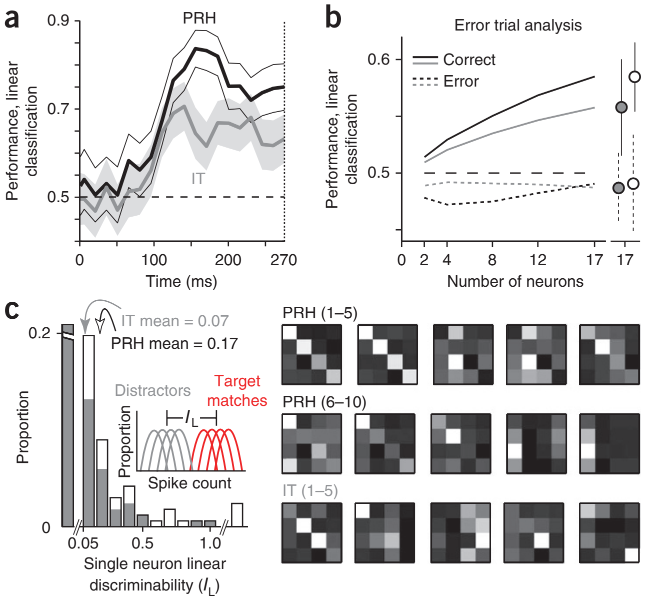
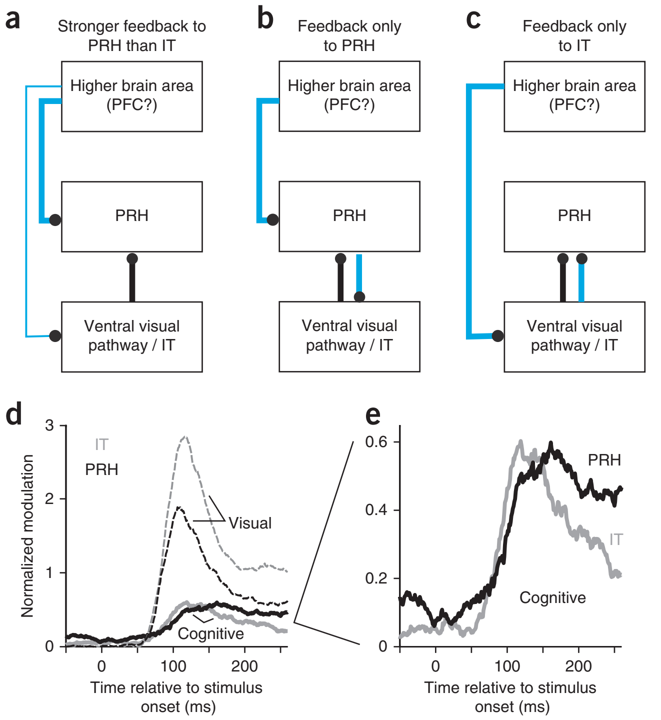


Figure 4 Additional population performance measures. (**a**) Evolution of linear classification performance over time. Thick lines indicate performance of the entire IT (gray) and PRH (black) populations for counting windows of 30 ms with 15-ms shifts between neighboring windows. Thin lines indicate standard error. The dashed line indicates the minimum reaction time on these trials (270 ms). (**b**) Linear classification performance on error (dotted) as compared with correct (solid) trials (data are presented as in **Fig. 3b**, left; see Online Methods). Each error trial was matched with a randomly selected correct trial that had the same target and visual stimulus as the condition that resulted in the error and both sets of trials were used to measure cross-validated performance when the population read-out was trained on separately measured correct trials, as described above. Error trials included both misses (of target matches) and false alarms (responding to a distractor). We performed the analysis separately for each multi-channel recording session and then averaged across sessions. Error bars, s.e.m. (**c**) Left, histograms of linearly separable target match information (I_L; equation (3), Online Methods), computed for IT (gray) and PRH (white). Arrows indicate means. The last bin includes PRH neurons with I_L of 1.1, 1.4, 3.3 and 5.3. The first (broken) bin includes IT and PRH neurons with negligible I_L (defined as $I_L < 0.05$; proportions = 0.75 in IT and 0.56 in PRH). Right, response matrices of the I_L top-ranked PRH and IT neurons (data are presented as in **Fig. 2b**) and the rankings labeled.

Figure 5 Discriminating between classes of models that predict more untangled target match information in PRH than IT. (**a**–**c**) Black lines indicate visual input and cyan lines indicate cognitive input, which can take the form of working memory or target match information. (**d**) Average magnitudes of visual (dashed) and cognitive (solid) normalized modulation plotted as a function of time relative to stimulus onset for IT (gray) and PRH (black). Normalized modulation was quantified as the bias-corrected ratio between signal variance and noise variance (equation (4), Online Methods), and provided a noise-corrected measure of the amount of neural response variability that could be attributed to: visual, changing the identity of the visual stimulus, or cognitive, changing the identity of the sought target and/or nonlinear interactions between changes in the visual stimulus and the sought target. (**e**) Enlarged view of the cognitive signals plotted in **d**. In **d** and **e**, response matrices were calculated from spikes in 60-ms bins with 1-ms shifts between bins.



a function of the separation of the responses to the target match and distractor conditions (equation (3) in Online Methods and **Fig. 4c**). This measure maps directly onto the amount of diagonal structure in a neuron's response matrix (Online Methods), and an idealized four-object target detector will therefore have high I_L, a single-object target detector will have a bit less, and a highly visual neuron or working memory neuron will have none (**Fig. 2b**). Consistent with the population results (**Fig. 3b**), we found that PRH had significantly higher mean single-neuron linearly separable target match information than IT ($P < 0.0001$; **Fig. 4c**). To relate our single neuron and population performance measures, we ranked the neurons in each population by their I_L and recomputed population performance as a function of the number of best neurons included in the population. In PRH, the best neurons were indeed four-object target detectors (**Fig. 4c**) and performance saturated fairly quickly as a function of population size (**Supplementary Fig. 2d**). In contrast, in IT, we found that the best neurons were detectors for at most two objects as targets (**Fig. 4c**) and IT performance was lower than PRH performance for equal-sized populations (**Supplementary Fig. 2d**). These results suggest that the compelling four-object target detectors that we found in PRH were responsible for a large portion of the population performance differences that we uncovered between IT and PRH. However, even after removing the best neurons (as many as 23) from PRH, performance in PRH remained higher than in IT (**Supplementary Fig. 2d**). Notably, many of the top 23 PRH neurons had single-object target detector structure (**Fig. 4c**). Together, these results suggest that higher PRH linear classifier performance can be attributed both to the existence of four-object target detectors that are absent in IT and to neurons with single-object target detector structure that are present in both areas, but are more numerous in PRH.

IT and PRH contain similar total target match information

Higher task performance in PRH versus IT when probed with a linear population read-out could reflect more total task-relevant information in PRH (because PRH receives task-relevant input that IT does not). Alternatively, these results could arise from a scenario in which IT and PRH contain similar amounts of total task-relevant information, but that information might be formatted such that it is less accessible to a linear read-out in IT than in PRH (**Fig. 3a**). To discern between these alternatives, we probed the total information for this task in a manner that did not depend on the specific format of that information. More specifically, total information for this task depended only on the degree to which the response clouds corresponding to target match and distractor conditions were non-overlapping, but not on the specific manner in which the response clouds were positioned relative to one another (**Fig. 3a**). As a measure of the total information available for match versus distractor discrimination in the IT and PRH populations, we performed a cross-validated, ideal observer match versus distractor classification of the population response on individual trials (equation (2), Online Methods).

We found that this measure of total task-relevant information was slightly lower in IT, but not significantly so ($P = 0.07$; **Fig. 3b**). Notably, even when the number of PRH neurons was halved relative to IT (50 PRH neurons versus 100 IT neurons), such that IT ideal observer performance was slightly higher than PRH (PRH = 86%, IT = 88%), linear classifier performance remained higher in PRH (PRH = 80%, IT = 66%). These results indicate that IT and PRH contain similar amounts of total information for this task, and that information is more tangled in IT and more untangled in PRH (**Fig. 3a**).

Evidence for feedforward untangling between IT and PRH

More untangled target match information in PRH as compared with IT could reflect a variety of mechanisms that differ in terms of the flow of information to and between IT and PRH. Here we consider three such general schemes. In each case, we refer to cognitive signals as the combination of all types of target-dependent modulation, including response modulations that can be attributed to changing the identity of the target and/or whether the stimulus was a match or a distractor. Notably, these schemes can be distinguished via their predictions about the relative amounts and/or the timing of cognitive information in IT as compared with PRH.

In the first scheme (**Fig. 5a**), cognitive information is fed back to both brain areas, and stronger PRH diagonal signals are accounted for by a stronger cognitive input to PRH as compared with IT. This class includes models in which cognitive information takes the form of a working memory input that is combined with visual information in IT and PRH, as well as models in which the diagonal signal is computed elsewhere and is then fed back to these two areas; in both cases, the magnitude of the combined cognitive modulation is predicted to be larger in PRH than in IT.

Second (**Fig. 5b**), stronger PRH diagonal signals may be accounted for by cognitive information that is fed back exclusively to PRH, which in turn passes some of this information back to IT. As in the first scheme, this cognitive information may take the form of a working memory and/or a diagonal signal. In either case, this scheme predicts that cognitive information should arrive earlier in PRH than in IT.

Third (**Fig. 5c**), cognitive information may be exclusively fed back to IT. Accounting for stronger diagonal signals with this scheme requires that cognitive signals are combined with visual signals in IT in a tangled manner, such that they are not accessible via a linear read-out, and that untangling computations in PRH reformat this information such that it becomes more linearly accessible. This class of models predicts that the magnitude of cognitive information should be approximately matched in the two brain areas and that cognitive information should arrive earlier in IT than in PRH.

To test the predictions of these three schemes, we performed a modified ANOVA analysis to parse each neuron's responses into firing rate modulations that could be attributed to changing the visual image, changing the cognitive context and noise resulting from trial-by-trial variability (equation (4), Online Methods). We found that cognitive modulations were approximately equal in strength in IT and PRH and that these modulations arrived slightly earlier in IT than in PRH (**Fig. 5d**,**e**), consistent with the third scheme, in which cognitive information is fed back only to IT and PRH inherits its cognitive information from IT as opposed to other sources (**Fig. 5c**; but see below). A decomposition of the combined cognitive signal into its linear (working memory) and nonlinear (interaction) components revealed that, consistent with other reports[22], working memory signals during the delay period (persistent activity) are present, but are weak in both areas (**Supplementary Fig. 3c**). In addition, the nonlinear component predominated during the stimulus-evoked response period (**Supplementary Fig. 3c**), consistent with either working memory signals that combine nonlinearly (for example, multiplicatively) with visual signals in these areas[23] or with visual and working memory combinations that are inherited from elsewhere (for example, V4).

Our results cannot definitively rule out some alternate proposals. For example, variants of a model in which IT and PRH both receive the same strength working memory input, but have different rules of combination (to produce tangled signals in IT and more untangled signals in PRH), would predict responses that are indistinguishable from the model we provide evidence for here (**Fig. 5c**). In addition, similar to other hierarchical descriptions of information processing[16,24,25], we do not know that PRH receives its information via a direct projection from IT to PRH (for example, information may first flow through the pulvinar or some other structure). In the next section, we evaluate the degree to which the class of functional models that are mathematically equivalent to the model proposed in **Figure 5c** can quantitatively account for our recorded responses. Similar to other functional model descriptions[23–28], the value of taking this type of approach is that it has the potential to provide insight into the algorithms by which information is transformed as it propagates through the brain (from IT to PRH), even in the absence of certainty regarding its exact biological implementation[29].

Taken together, these results (**Figs. 3**–**5**) are consistent with a functional model in which visual and working memory signals are initially combined in or before IT in the ventral visual pathway in a heterogeneous and tangled manner, followed by reformatting operations in PRH that untangle target match information. These results are reminiscent of the untangling phenomena described at earlier stages of the ventral visual pathway (from V1 to V4 to IT) for invariant object recognition[16,30–32], and therefore suggest that the brain transforms information into a format that can be accessed via a linear population read-out not only for perception (identifying the content of a currently-viewed scene), but also for more cognitive tasks (finding a specific sought target object).

A linear-nonlinear model can account for PRH untangling

Next, we were interested in whether an untangling transformation from IT to PRH could provide an accurate quantitative account of our data. Thus, we set out to determine the simplest class of models that could take our recorded IT responses as input and produce a model population that had properties similar to our recorded PRH. We immediately ruled out the class of models in which IT neurons combine linearly to produce PRH cells, as we know that linear operations can move linearly separable information around in a population (that is, between neurons), but cannot transform nonlinearly separable information into a linearly separable format. Thus, we began by testing the class of nonlinear models in which a static nonlinearity (thresholding and saturation; equation (10) in Online Methods) was fit to each IT neuron such that its response matrix conveyed maximal linearly separable target match information (I_L; **Fig. 4c**). Inconsistent with the large gains in linear read-out performance we observed from IT to PRH, we found only modest overall gains in this model population (**Fig. 6a**).

We then considered the class of models in which pairs of IT neurons combine via a linear-nonlinear model to produce the responses of pairs of PRH cells (**Fig. 6b**). In fitting our model, we imposed the constraint that information could not be replicated multiple times in the transformation from IT to PRH (that is, the same neuron could not be copied multiple times). To enforce this rule, our model created two PRH neurons by applying two sets of orthonormal linear weights to

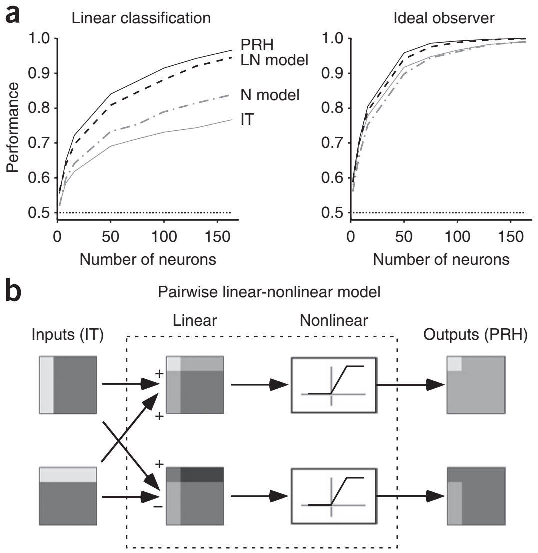


Figure 6 Modeling the transformation from IT to PRH. (**a**) Shown are linear classification (left) and ideal observer (right) performance of the following populations: IT (gray), PRH (black), the nonlinear (N) model (gray dot-dashed) and the linear-nonlinear (LN) model (black dashed). Data are presented as in **Figure 3b**. To compare performance of the actual and model populations, we regenerated Poisson trial-by-trial variability for the actual IT and PRH populations from the mean firing rate responses across trials (the response matrix) for each IT and PRH neuron. (**b**) The pairwise linear-nonlinear model we fit to describe the transformation from IT to PRH, shown for two idealized IT neurons. To create the linear-nonlinear model, we combined pairs of IT neurons via two sets of orthogonal linear weights, followed by a nonlinearity to create two model PRH neurons.

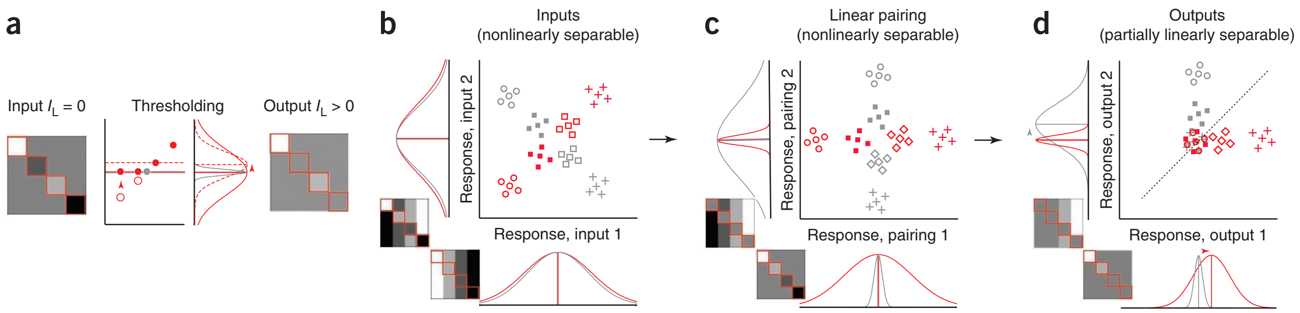


Figure 7 The neural mechanisms underlying untangling. (**a**) Shown is an idealized neuron that has the same average response to matches (red solid) and distractors (gray), and thus no linearly separable information (I_L = 0). However, because the lowest responses in the matrix are matches (red open circles), a threshold nonlinearity can set these to a higher value (red solid circles), thereby producing an increase in the overall mean match response (red dashed) such that it is now higher than the average distractor response (gray). Because linearly separable information depends on the difference between these means, this translates directly into an increase in linearly separable information in the output neuron (I_L > 0). (**b**) Two idealized neurons presented as in **Figure 2b**. The two neurons produce a nonlinearly separable representation in which a linear decision boundary is largely incapable of separating matches from distractors. However, these two idealized neurons have perfect tuning correlations for matches and perfect tuning anti-correlations for distractors. (**c**) Pairing the two neurons via two sets of orthogonal linear weights produces a rotation in the two-dimensional space and a difference in the response variance for matches and distractors for both neurons. (**d**) Applying a nonlinearity to the linearly paired responses results in a representation in which a linear decision boundary is partially capable at distinguishing matches and distractors. The effectiveness of pairing can be attributed to an asymmetry (that is, a difference) in the neurons' tuning correlations for matches and distractors (equation (24), Online Methods).

the pair of IT inputs (for example, $\left(+\sqrt{0.5},+\sqrt{0.5}\right)$ and $\left(+\sqrt{0.5},-\sqrt{0.5}\right)$) and each IT neuron was included only once (equations (12–14), Online Methods). We searched all possible pairwise combinations of IT neurons and nonlinearities and selected the combinations that produced the largest gains in linearly separable information (Online Methods). The resulting linear-nonlinear model population nearly matched the population performance increases in PRH over IT with a linear read-out and replicated PRH population performance on the match versus distractor task with an ideal observer read-out (**Fig. 6a**). The linear-nonlinear model also replicated a number of single-neuron response differences in PRH relative to IT, including a decrease in the visual modulation strength and an increase in the congruency (the alignment) of visual and target signals (**Supplementary Fig. 4**), despite the fact that the model was not explicitly fit to account for these parameters. The fact that such a simple model reproduced the transformation that we observed in our data from IT to PRH provides support for the proposal that PRH receives its inputs for this task primarily from IT, as opposed to other sources. The simplicity of the model also lends itself to an exploration of the specific computational mechanisms underlying untangling.

Untangling relies on asymmetric tuning correlations in IT

To understand how the pairwise linear-nonlinear model untangles information, it is useful to first conceptualize how a nonlinearity can act to increase linearly separable information (I_L) in a neuron's matrix. A nonlinearity can be effective in situations in which the variance (the spread) across one set of conditions (for example, the matches) is higher than the other set (for example, the distractors) (**Fig. 7a**). In such scenarios, the nonlinearity can change a subset of responses in the high variance set and therefore increase the difference between the mean response to matches and distractors (**Fig. 7a); this** translates into an increase of the amount of linearly separable target match information (equation (19), Online Methods). Our results (**Fig. 6a**) suggest that pairing is important for producing linearly separable information (as compared with applying a nonlinearity without pairing). How does pairing make a nonlinearity more effective? We can envision the responses of these neurons in a population space similar to that depicted in **Figure 3a** (a population of size 2), where the representation of target matches and distractors is initially nonlinearly separable or tangled (**Fig. 7b**). In this example, the firing rate distributions of both neurons have the same mean response to matches and distractors (hence no linearly separable information) and the same variance in their responses to matches and distractors (hence a nonlinearity applied to either of them would produce no increase in linearly separable information). However, a rotation of the population response space produces a variance difference between matches and distractors for both neurons (**Fig. 7c**), and, thus, a scenario in which a nonlinearity is effective at producing a more linearly separable representation (**Fig. 7d**). This type of rotation can be achieved by pairing the two neurons via orthogonal linear weights (for example, positive weights for one pairing and a positive and negative weight for the other pairing). In general, a linear pairing of two neurons tends to be effective when the two neurons have asymmetric tuning correlations for matches and distractors (for example, a positive correlation, or similar tuning, for matches and a negative correlation, or the opposite tuning, for distractors). When two such neurons are combined, these tuning correlation asymmetries translate into variance differences between matches and distractors, yielding a scenario in which a nonlinearity will be effective at producing a representation that can be better accessed via a linear read-out (**Fig. 7c,d**).

We formalized these intuitions into a quantitative prediction of the amount of linearly separable information that can be gained by pairing any two IT neurons via a linear-nonlinear model of the form we fit to our data; our prediction relies on the degree of asymmetry in the neurons' match and distractor tuning correlations (equations (22 and 24), Online Methods). Empirically, we found that this prediction provided a good account of the linearly separable information extracted by our linear-nonlinear model of the transformation from IT to PRH (correlation of the actual and predicted information gains for each pair, r = 0.84), confirming that the asymmetric tuning correlation mechanism is a good description of how the pairwise linear-nonlinear model untangles information.

This description of untangling via asymmetric tuning correlations reveals that, for any given IT neuron, its best possible pair is one that has a perfect tuning correlation for one set (for example, matches) and a perfect tuning anti-correlation for the other set (for example, distractors). However, we note that modest tuning correlation asymmetries are also predicted to translate into increases in linearly separable information (under appropriate conditions; equation (19), Online Methods). We found that our model did largely rely on modest (as opposed to maximal) tuning correlation asymmetries (**Supplementary Fig. 5a**–**d**) and that such modest tuning correlation asymmetries are ubiquitously present in populations of neurons that reflect mixtures of visual and target signals (**Supplementary Fig. 5e**,**f**).

DISCUSSION

Finding specific targets requires the combination of visual and target-specific working memory signals. The ability to flexibly switch between different targets imposes the computational constraint that this combination must be followed by a reformatting process to construct a signal that reports whether a target is present in a currently viewed scene (**Fig. 1**). Although the locus of the combination of visual and target-specific signals is thought to reside at mid-to-higher stages of the ventral visual pathway[7–13], the rules by which the brain combines and reformats this information are not well understood. Our results build on earlier studies to discriminate between models that describe where and how visual and target signals combine (**Fig. 5**), provide a functional model in which visual and target-specific signals combine to produce a linearly inseparable or tangled representation of target matches in IT that is then untangled in PRH (**Figs. 3**, **4** and **6** and **Supplementary Fig. 4**), and provide a neural mechanism that can account for the untangling or reformatting process (**Fig. 7**, **Supplementary Fig. 5**). Notably, our results are not predictable from earlier reports. Specifically, a series of studies reported signals that differentiate target matches from distractors not only in PRH[15,33], but also in V4 (ref. 8) and IT[11]. Thus, it has been difficult to discern the degree to which the target match signals present in PRH are inherited from combinations of visual and working memory inputs at earlier stages of the ventral visual pathway (for example, V4 and IT), as compared with working memory inputs directly to PRH. Although we cannot definitively rule out the latter hypothesis, our results indicate that, consistent with the former suggestion, the task-specific information contained in PRH is also present in an earlier structure (but contained in a different format). Moreover, our results provide both a computational (untangling) and mechanistic (pairing via asymmetric tuning correlations) description of how that information might be reformatted in a feedforward scheme.

Although not definitive, a number of lines of evidence support a model in which PRH reformats information arriving (directly or indirectly) from IT. First, anatomical evidence suggests that the primary input to PRH is in fact IT[17]. Second, our results indicate that nearly all of the information for this task found in PRH was also contained in IT, suggesting that PRH need not get its input from other sources (**Fig. 3b**). Third, the relative amounts and timing of cognitive signals were consistent with this description (**Fig. 5e**). Finally, our results suggest that a simple linear-nonlinear model can account for the transformation (**Fig. 6a**, **Supplementary Fig. 4**). As described above, ours is a functional model of neural computation that describes how signals are transformed as they propagate from one stage of processing (IT) to a higher brain area (PRH). Similar to other functional model descriptions[23–28], we cannot rule out alternate proposals that predict the same neural responses, but have different pathways (for example, additional structures or parallel inputs) for the flow of information.

Our results reveal that visual and working memory signals are combined in a manner that results in a largely tangled representation of target-match information in IT. This finding is consistent with visual and working memory signals that are combined, in part, via misaligned or incongruent object preferences (for example, to produce the distractor detector in **Fig. 2b**; see also **Supplementary Fig. 4**). Similar incongruent neurons have been reported elsewhere[8,11]. If the brain could (in theory) achieve an untangled representation at the locus of combination by congruently combining visual and working memory signals, why might it instead combine these signals in a tangled and partially incongruent fashion only to untangle them downstream? We do not know, but we can speculate. First, working memory signals corresponding to a sought target are likely to be fed back to higher stages of the visual system (from PFC to V4 and/or IT) and because V4 and IT lack a precise topography for object identity, developing circuits that precisely align these two types of signals may be challenging[34]. Second, having signals that report incongruent combinations might be functionally advantageous for tasks that are more complex than the one we present here[35]. For example, incongruent signals might be useful during visual search tasks when evaluating where to look next (for example, "I am looking for my car keys and I am looking at my wallet; my keys are likely to be nearby"[36]).

Our results describe a mechanism by which information may be reformatted in PRH by combining IT neurons with asymmetric tuning correlations. Similar to other functional models[23,24,26–28,37], our model is designed to capture neural computation in a simplified manner that is not directly biophysical, but can be mapped onto biophysical mechanism. How might untangling via linear-nonlinear pairings of neurons with asymmetric tuning correlations be implemented in the brain? Although simple pairwise combinations of IT neurons were sufficient to explain the responses that we observed in PRH, each input probably reflects a functional pool of hundreds of neurons that (directly or indirectly) project from IT to a particular site in PRH[17]. Such connections could be wired via a reinforcement learning algorithm (for example, ref. 38) during the natural experience of searching for targets.

Our results demonstrate that target match information is formatted in a manner more accessible to a simple, linear read-out in PRH than in IT. Although we do not know the precise rules that the brain uses to read-out target match information, mechanistically, we envision that this could be implemented in the brain by a higher order neuron that 'looks down' on a population and determines whether a target is in view. Simple decision boundaries, such as linear hyperplanes, are consistent with the machinery that can be implemented by an individual neuron (for example, a weighted sum of its inputs, followed by a threshold), whereas highly nonlinear decision boundaries are likely beyond the computational capacity of neurons at a single stage[16,32]. Does PRH reflect a fully untangled representation of target match information? Probably not. Although other studies have also suggested that the responses of PRH neurons explicitly reflect target match information[15,33], PFC neurons have been reported to convey more target match information than neurons in PRH[5]. Given that PRH projects to PFC[39], the representation of target matches reflected in PRH may be further untangled in PFC and used to guide behavior. Alternatively, target match information reflected in PRH and PFC might constitute different pathways (for example, from PRH, signals might propagate more deeply into the temporal lobe) and might be used for different purposes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS
We thank Y. Cohen, J.A. Movshon, E. Simoncelli and A. Stocker for helpful discussions. We are especially grateful to D. Brainard and J. Gold for detailed

comments on the work and on the manuscript. We also thank J. Deutsch for technical assistance and C. Veeder for veterinary support. This work was supported by the National Eye Institute of the US National Institutes of Health (award number R01EY020851), a Sloan Foundation award to N.C.R. and contributions from a National Eye Institute core grant (award number P30EY001583).

AUTHOR CONTRIBUTIONS

N.C.R., M.P.W. and M.P. conducted the experiments. M.P. and L.S.U. developed the data alignment software. M.P.W. and N.C.R. sorted the spike waveforms. M.P. and N.C.R. developed and executed the analyses. M.P. and N.C.R. wrote the manuscript. N.C.R. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Salinas, E. Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *J. Neurosci.* **24**, 1113–1118 (2004).
2. Salinas, E. & Bentley, N.M. Gain modulation as a mechanism for switching reference frames, tasks and targets. in *Coherent Behavior in Neuronal Networks* (eds. Josic, K., Rubin, J., Matias, M. & Romo, R.) 121–142 (Springer, New York, 2009).
3. Engel, T.A. & Wang, X.J. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *J. Neurosci.* **31**, 6982–6996 (2011).
4. Sugase-Miyamoto, Y., Liu, Z., Wiener, M.C., Optican, L.M. & Richmond, B.J. Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Comput. Biol.* **4**, e1000073 (2008).
5. Miller, E.K., Erickson, C.A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
6. Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I. & Miyashita, Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401**, 699–703 (1999).
7. Haenny, P.E., Maunsell, J.H.R. & Schiller, P.H. State-dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exp. Brain Res.* **69**, 245–259 (1988).
8. Maunsell, J.H.R., Sclar, G., Nealey, T.A. & Depriest, D.D. Extraretinal representations in area V4 in the macaque monkey. *Vis. Neurosci.* **7**, 561–573 (1991).
9. Bichot, N.P., Rossi, A.F. & Desimone, R. Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* **308**, 529–534 (2005).
10. Chelazzi, L., Miller, E.K., Duncan, J. & Desimone, R. Responses of neurons in macaque area V4 during memory-guided visual search. *Cereb. Cortex* **11**, 761–772 (2001).
11. Eskandar, E.N., Richmond, B.J. & Optican, L.M. Role of inferior temporal neurons in visual memory. II. Temporal encoding of information about visual images, recalled images and behavioral context. *J. Neurophysiol.* **68**, 1277–1295 (1992).
12. Liu, Z. & Richmond, B.J. Response differences in monkey TE and perirhinal cortex: stimulus association related to reward schedules. *J. Neurophysiol.* **83**, 1677–1692 (2000).
13. Gibson, J.R. & Maunsell, J.H.R. Sensory modality specificity of neural activity related to memory in visual cortex. *J. Neurophysiol.* **78**, 1263–1275 (1997).
14. Lueschow, A., Miller, E.K. & Desimone, R. Inferior temporal mechanisms for invariant object recognition. *Cereb. Cortex* **4**, 523–531 (1994).
15. Miller, E.K. & Desimone, R. Parallel neuronal mechanisms for short-term memory. *Science* **263**, 520–522 (1994).
16. DiCarlo, J.J. & Cox, D.D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
17. Suzuki, W.A. & Amaral, D.G. Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *J. Comp. Neurol.* **350**, 497–533 (1994).
18. Meunier, M., Bachevalier, J., Mishkin, M. & Murray, E.A. Effects on visual recognition of combined and separate ablations of the entorhinal and perirhinal cortex in rhesus monkeys. *J. Neurosci.* **13**, 5418–5432 (1993).
19. Buffalo, E.A., Ramus, S.J., Squire, L.R. & Zola, S.M. Perception and recognition memory in monkeys following lesions of area TE and perirhinal cortex. *Learn. Mem.* **7**, 375–382 (2000).
20. Cohen, M.R. & Maunsell, J.H. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594–1600 (2009).
21. Graf, A.B., Kohn, A., Jazayeri, M. & Movshon, J.A. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat. Neurosci.* **14**, 239–245 (2011).
22. Fuster, J.M. & Jervey, J.P. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* **212**, 952–955 (1981).
23. Reynolds, J.H. & Heeger, D.J. The normalization model of attention. *Neuron* **61**, 168–185 (2009).
24. Rust, N.C., Mante, V., Simoncelli, E.P. & Movshon, J.A. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* **9**, 1421–1431 (2006).
25. Gold, J.I. & Shadlen, M.N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions and reward. *Neuron* **36**, 299–308 (2002).
26. Simoncelli, E.P. & Heeger, D.J. A model of neuronal responses in visual area MT. *Vision Res.* **38**, 743–761 (1998).
27. Heeger, D.J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
28. Adelson, E.H. & Bergen, J.R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
29. Marr, D. *Vision* (MIT Press, Cambridge, Massachusetts, 1982).
30. Rust, N.C. & DiCarlo, J.J. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
31. Hung, C.P., Kreiman, G., Poggio, T. & DiCarlo, J.J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
32. DiCarlo, J.J., Zoccolan, D. & Rust, N.C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
33. Chelazzi, L., Miller, E.K., Duncan, J. & Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347 (1993).
34. Maunsell, J.H.R. & Treue, S. Feature-based attention in visual cortex. *Trends Neurosci.* **29**, 317–322 (2006).
35. Rigotti, M., Ben Dayan Rubin, D., Wang, X.J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front Comput. Neurosci.* **4**, 24 (2010).
36. Najemnik, J. & Geisler, W.S. Optimal eye movement strategies in visual search. *Nature* **434**, 387–391 (2005).
37. Shadlen, M.N. & Newsome, W.T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* **86**, 1916–1936 (2001).
38. Law, C.T. & Gold, J.I. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat. Neurosci.* **12**, 655–663 (2009).
39. Lavenex, P., Suzuki, W.A. & Amaral, D.G. Perirhinal and parahippocampal cortices of the macaque monkey: projections to the neocortex. *J. Comp. Neurol.* **447**, 394–420 (2002).
40. Minsky, M. & Papert, S. *Perceptrons: an Introduction to Computational Geometry* (MIT Press, Cambridge, Massachusetts, 1969).

ONLINE METHODS

The subjects in this experiment were two naive adult male rhesus macaque monkeys (8.0 and 15.0 kg). Aseptic surgeries were performed to implant head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli, reward and data acquisition were controlled using customized software (http://mworks-project.org). Stimuli were presented on a LCD monitor with a 85-Hz refresh (Samsung 2233RZ)[41]. Both IT and PRH were accessed via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images and later verified physiologically by the locations and depths of gray and white matter transitions that included characteristic transitions through subcortical structures (for example, the putamen and amygdala) to reach PRH. The region of IT recorded was located on both the ventral superior temporal sulcus and the ventral surface of the brain, over a 4-mm medial-lateral region located lateral to the anterior middle temporal sulcus that spanned 14–17 mm anterior to the ear canals[12,30]. The region of PRH recorded was located medial to the anterior middle temporal sulcus and lateral to the rhinal sulcus and extended over a 3-mm medial-lateral region located 19–22 mm anterior to the ear canals[12]. We recorded neural activity via a combination of glass-coated tungsten single electrodes (Alpha Omega) and 16- and 24-channel U probes with recording sites arranged linearly and separated by 150 μm spacing (Plexon). Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored via the OmniPlex Data Acquisition System (Plexon). We performed all spike sorting manually offline using commercially available software (Plexon). Although we were not blind to the brain area recorded in each session, we attempted to record from any neural signals that we could isolate within the predefined brain areas irrespective of their response properties and we did not perform any online data analyses to select specific recording locations. In addition, our offline spike sorting procedures were performed blind to the specific experimental conditions (whether a condition was a target match or a distractor) and our data analyses were automated to avoid the introduction of bias. The number of neurons that we recorded (our sample size) was designed to approximately match previous publications[30]; no statistical tests were run to determine the sample size a priori. Monkeys initiated a trial by fixating a small dot. After a 250-ms delay, an image indicating the target was presented, followed by a random number (0–3, uniformly distributed) of distractors and then the target match. Each image was presented for 400 ms, followed by a 400-ms blank. Monkeys were required to maintain fixation throughout the distractors and make a saccade to a response dot located 7.5 degrees below fixation after 150 ms following target onset, but before the onset of the next stimulus to receive a reward. The same four images were used during all the experiments. Approximately 25% of trials included the repeated presentation of the same distractor with zero or one intervening distractors of a different identity. The same target remained fixed in short blocks of ~1.7 min that included an average of nine correct trials. In each block, four presentations of each condition (for a fixed target) were collected and all four target blocks were presented in a 'metablock' in pseudorandom order before reshuffling. A minimum of five metablocks in total (20 correct presentations for each experimental condition) were collected.

Responses were only analyzed on correct trials, unless otherwise stated. Target matches that were presented after the maximal number of distractors ($n = 3$) occurred with 100% probability and were discarded from the analysis. Unless otherwise stated, we measured the response of each neuron as the spike count in a time window 80–270 ms after stimulus onset. To maximize the length of our counting window, but also ensure that spikes were only counted during periods of fixation, we randomly selected responses to target matches from the 74.2% of correct trials on which the monkeys' reaction times exceeded 270 ms. Including trials with faster reaction times did not change the results (that is, claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys; **Supplementary Fig. 2c**). As a measure of unit isolation, we determined the signal-to-noise ratio (SNR) of each spike waveform as the difference between the maximum and minimum of the mean waveform trace, divided by twice the s.d. across the differences between the actual waveforms and the mean waveform[42]. We screened units by their SNR and by a one-way ANOVA to determine those units whose firing rates were significantly modulated by the task parameters. When determining the screening criteria to include units in our analysis, we were concerned that setting any particular fixed value, particularly a highly stringent value, might differentially affect the two populations (for example, as a result of lower firing rates in one of our populations). The most liberal screening procedure that we applied (one-way ANOVA, $P < 0.05$ and SNR > 2) resulted in 167 and 164 units in IT and PRH, respectively, and for all but the analysis shown in **Figure 4b** and **Supplementary Figure 2b**, these are the criteria we used. The SNR was not statistically different in the two resulting populations, as assessed by a statistical comparison of their means (mean IT = 3.47, PRH = 3.55, $P = 0.55$). Applying increasingly stringent criteria to the ANOVA (to $P < 0.0001$) or to unit isolation (to SNR > 3.5) did not change the results (that is, claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys).

To assess the effect of simultaneous trial-by-trial variability (that is, noise correlations) on population performance (**Supplementary Fig. 2b**), we analyzed data simultaneously collected on the multi-channel U probes (described above). During spike sorting, we defined at least one unit on every available channel and we determined the 17 units from each session that produced the most significant P values in the one-way ANOVA screen (without setting an absolute threshold on this P value nor on SNR isolation). We assessed linear classifier performance for these simultaneously recorded subpopulations in the manner described below. We used a similar approach to compute population performance on error trials (**Fig. 4b**). Specifically, for each multi-channel recording session, we determined misses as instances in which the monkey failed to break fixation in response to the target match and false alarms as instances in which the monkey's eyes made a downward saccade in response to a distractor. We confined our analysis to false alarms in which the monkey fixated for a minimum of 270 ms before the response and for both types of error trials, we counted spikes in the same window used on correct trials (80–270 ms after stimulus onset). We compared linear classifier performance on error and correct control trials in the manner described below.

Population performance. To determine population measures of the amount and format of information available in IT and PRH to discriminate target matches and distractors, we performed a series of classification analyses. Specifically, we considered the spike count responses of a population of N neurons to P presentations of M images as an N-dimensional population response vector **x**. We performed a series of cross-validated procedures in which (unless otherwise stated) we randomly assigned 80% of our trials (16 trials) to compute the representation (training trials) and we set aside the remaining 20% of our data (four trials) to test the representation (test trials).

Linear classification. To determine how well each population could discriminate target matches from distractors across changes in target identity using a linear decision rule, we implemented a linear readout procedure similar to that used in previous studies[30]. The linear readout amounted to using the training data to find a linear hyperplane that would best separate the population response vectors corresponding to all of the target match conditions from the response vectors corresponding to distractors (**Fig. 3b**). The linear readout took the form

$$f(x) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \qquad (1)$$

where **w** is an N-dimensional vector describing the linear weight applied to each neuron (and thus defines the orientation of the hyperplane), and b is a scalar value that offsets the hyperplane from the origin and acts as a threshold. The population classification of a test response vector was assigned to a target match when $f(x)$ exceeded zero and was classified as a distractor otherwise. The hyperplane and threshold for each classifier were determined by a support vector machine procedure using the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) with a linear kernel, the C-SVC algorithm, and cost (C) set to 0.1.

Ideal observer classification. To determine how well each population could discriminate target matches from distractors across changes in target identity using an ideal observer, we computed from the training trials the average spike count response r_{uc} of each neuron u to each of the 16 different conditions c. Assuming Poisson trial-by-trial variability, the likelihood that a test response k arose from a particular condition for a neuron was computed as the Poisson probability density

$$lik_{u,c} = \frac{(r_{uc})^{k} \cdot e^{-r_{uc}}}{k!} \qquad (2)$$

We then computed the likelihood that a test response vector **x** arose from each condition c for the population as the product of the likelihoods for the individual neurons. Finally, we computed the likelihood that a test response vector arose from the category target match versus the category distractor as the mean of the likelihoods for target matches and distractors, respectively. The population classification was assigned to the category with the higher likelihood (**Fig. 3b**).

To compare population performance between the different classifiers, we performed the same resampling procedure for each of them. On each iteration of the resampling, we randomly assigned trials without replacement for training and testing and when subpopulations with fewer than the full population were tested, we randomly selected a new subpopulation of neurons without replacement from all neurons. Because some of our neurons were recorded simultaneously, but most of them were recorded in different sessions, unless otherwise stated, trials were shuffled on each iteration to destroy any (real or artificial) trial-by-trial correlation structure that might exist between neurons. Our experimental design resulted in four target match conditions and 12 distractor conditions; on each iteration we randomly selected one distractor condition from each image (for a total of four distractor conditions) to avoid artificial overestimations of classifier performance that could be produced by taking the prior distribution into account (for example, scenarios in which the answer is more likely to be distractor than target match). We calculated means and standard error for performance as the mean and s.d., respectively, across 200 resampling iterations. To evaluate the probability that differences in mean performance between IT and PRH were due to chance, we calculated *P* values as the fraction of the 200 bootstrap iterations in which the lower mean exceeded the higher mean, for populations with the maximum number of neurons.

To assess the effect of correlated noise on population performance, we compared classifier performance when the trial-by-trial variability was kept intact as compared to when it was randomly shuffled (**Supplementary Fig. 2b**) for populations of 17 simultaneously recorded sites (where the data were extracted in the manner described above). Performance was computed as the mean across recording sessions; standard error was computed as the s.d. across 200 iterations in which trials were randomly assigned as training and testing, and, for populations smaller than 17, the subset of neurons was randomly selected, and, for the shuffled noise case, trials were randomly shuffled. To compare performance on correct and error trials (**Fig. 4b**), we extracted the error trials from these same multi-channel recording sessions. For each error trial (misses and false alarms; described above), we randomly selected a correct trial condition that was matched for the same target and visual stimulus as the condition that led to the error. We set aside these correct (and error) trials for cross validation, and trained the linear classifier on separate correct trials, as described above. Performance on each resampling iteration was computed as the average across all recording sessions; standard error was computed as the s.d. across 800 resampling iterations in which correct trials were randomly assigned as training and test, and, for populations smaller than 17, the subset of neurons were randomly selected.

Single-neuron measure of linearly separable target match information. As a single-neuron measure of match and distractor linear discriminability, we computed how well a neuron could linearly separate the responses to four target matches from the responses to 12 distractors (**Fig. 4c**). This was measured by the squared difference between the mean response to all target matches μ_{Match} and the mean response to all distractors $\mu_{\text{Distractor}}$ divided by the variance of the spike count across trials, averaged across all 16 conditions σ^2_{noise} (ref. 43)

$$I_{\text{L}} = \frac{(\mu_{\text{Match}} - \mu_{\text{Distractor}})^2}{\sigma^2_{\text{noise}}} \qquad (3)$$

Single-neuron measures of visual and cognitive information. We began by performing a two-way ANOVA to parse each neuron's total response variability σ^2_{tot} (that is, total variance across all trials and conditions) into four terms: modulation across visual stimuli σ^2_{vis}, modulation across sought targets σ^2_{targ}, nonlinear interactions of visual and target modulations σ^2_{NL}, and trial-by-trial variability σ^2_{noise}

$$\sigma^2_{\text{tot}} \cdot v_{\text{tot}} = \sigma^2_{\text{vis}} \cdot v_{\text{vis}} + \sigma^2_{\text{targ}} \cdot v_{\text{targ}} + \sigma^2_{\text{NL}} \cdot v_{\text{NL}} + \sigma^2_{\text{noise}} \cdot v_{\text{noise}} \qquad (4)$$

where $v_{\text{tot}} = 319$ (total number of degrees of freedom), $v_{\text{vis}} = 3$ (degrees of freedom of visual modulation), $v_{\text{targ}} = 3$ (degrees of freedom of target modulation), $v_{\text{NL}} = 9$ (degrees of freedom of visual and target modulation interactions) and $v_{\text{noise}} = 304$ (degrees of freedom of noise variability). We then computed the ratios of signal modulations and noise variability to establish the normalized magnitudes of visual, linear cognitive, nonlinear cognitive and total cognitive modulation. In particular, we calculated the fraction of a neuron's variance that could be attributed to changes in the identity of the visual image (**Fig. 5d** and **Supplementary Figs. 3** and **4**), normalized by the noise variability, as $\sigma^2_{\text{vis}} / \sigma^2_{\text{noise}}$ The fraction of a neuron's variance that could be attributed to changes in the target (that is, working memory signal; **Supplementary Fig. 3**) was captured by the variance of linear target modulations, normalized by the noise variability $\sigma^2_{\text{targ}} / \sigma^2_{\text{noise}}$. The fraction of a neuron's variance that could be attributed to nonlinear cognitive modulation (**Supplementary Fig. 3**) was captured by the variance of nonlinear interactions of visual and target identity, normalized by the noise variability $\sigma^2_{\text{NL}} / \sigma^2_{\text{noise}}$. The fraction of a neuron's variance that could be attributed to overall changes in the cognitive context (that is, overall cognitive signal; **Fig. 5d**,**e** and **Supplementary Fig. 4**) was captured by the combined variance that could be attributed to linear and nonlinear target modulations, normalized by the noise variability $\left(\sigma^2_{\text{targ}} + \sigma^2_{\text{NL}}\right) / \sigma^2_{\text{noise}}$.

Measuring the amount of signal modulation in the presence of noise and with a limited number of samples leads to an overestimation of the signal. For example, consider a hypothetical neuron that produces the exact same firing rate response to all task conditions; due to trial-by-trial variability, the computed average firing rate responses across trials will differ, thereby giving one the impression that the neuron does in fact respond differentially to the stimuli. To correct for this bias, we first estimated the amount of measured signal modulation that is expected under the assumption of zero 'true' signal: assuming Poisson variability, the bias is almost exactly equal to the number of degrees of freedom of the signal divided by the number of trials: bias $\approx v_{\text{signal}} / n$. Unbiased estimates were then obtained by subtracting this value from our information measurements.

Congruency. For those neurons that were significantly modulated (F test, $P < 0.05$) by both visual and target information or their interaction, we were interested in measuring the degree to which visual and target signals had been combined congruently (that is, with similar object preferences). In doing so, it became necessary to evaluate congruency for the linear (σ^2_{vis} and σ^2_{targ}) and nonlinear interaction (σ^2_{NL}) terms separately. We defined linear congruency as the absolute value of the Pearson correlation between the visual marginal tuning (that is, the average response to each image as the visual stimulus) and the target marginal tuning (that is, the average response to each image as the target)

$$\text{linear congruency} = \left|\rho\left(x_{\text{vis}}, x_{\text{targ}}\right)\right| \qquad (5)$$

$$x_{\text{vis}}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R\left(\text{vis} = i, \text{targ} = k\right),\ x_{\text{targ}}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R(\text{vis} = k, \text{targ} = i)$$

where $R\left(\text{vis} = i, \text{targ} = k\right)$ is the average response to visual stimulus i while searching for target k. To measure nonlinear congruency, we considered the nonlinear modulation σ^2_{NL} described above and we sought to determine the degree to which these modulations fell along the diagonal (that is, congruent nonlinear combinations of visual and target signals) versus off the diagonal (that is, incongruent combinations). We quantified this by parsing the total nonlinear variability σ^2_{NL} into a term capturing the diagonal modulation σ^2_{diag} and a term capturing the non-diagonal modulation $\sigma^2_{\text{nondiag}}$

$$\sigma^2_{\text{diag}} = \left(\mu_{\text{Match}} - \mu_{\text{Distractor}}\right)^2 / 3 \qquad (6)$$

$$\sigma^2_{\text{nondiag}} \cdot v_{\text{nondiag}} = \sigma^2_{\text{NL}} \cdot v_{\text{NL}} - \sigma^2_{\text{diag}} \cdot v_{\text{diag}}$$

where $v_{\text{NL}} = 9$ (degrees of freedom of nonlinear interactions, as above), $v_{\text{diag}} = 1$ (degrees of freedom of diagonal modulation) and $v_{\text{nondiag}} = 8$ (degrees of freedom of nondiagonal modulation). We defined nonlinear congruency as the ratio between diagonal modulation and the sum of diagonal and nondiagonal modulation

$$\text{nonlinear congruency} = \frac{\sigma^2_{\text{diag}}}{\sigma^2_{\text{diag}} + \sigma^2_{\text{nondiag}}} \qquad (7)$$

The final congruency index was computed as a weighted average of linear and nonlinear congruency, where the weights were determined by the firing rate variance for each term

$$I_{\text{congr}} = \frac{\sigma_{\text{lin}}^2 \cdot \text{linear congruency} + \sigma_{\text{NL}}^2 \cdot \text{nonlinear congruency}}{\sigma_{\text{lin}}^2 + \sigma_{\text{NL}}^2} \quad \text{where } \sigma_{\text{lin}}^2 = \sigma_{\text{vis}}^2 + \sigma_{\text{targ}}^2 \tag{8}$$

We designed the congruency index to range from 0 to 1 and to take on a value of 0.5 (on average) for random alignments of visual and working memory signals. Because the range of obtainable congruencies depends on a neurons tuning bandwidth and overall firing rate, as benchmarks for these values, we determined the upper and lower congruencies that could be achieved for each neuron by computing congruencies for all possible shuffles of its rows and columns; we found that the obtainable range was on average very broad (average minimum = 0.09, average maximum = 0.87) and we confirmed that random alignments of the rows and columns produce average congruency values near 0.5 (0.48).

Static nonlinear model of the transformation from IT to PRH. Our goal was to determine the class of models that could transform the responses of IT neurons into a new artificial neural population with the response properties that we observed in PRH (including increases in the amounts of untangled target match information). We fit the newly generated neurons to maximize the total amount of linearly separable target match information in the model population (I_L, see equation (3)). In our model neurons, we imposed Poisson trial-by-trial variability. We could therefore compute I_L by replacing the noise variance term σ_{noise}^2 with the mean responses across all conditions, μ

$$I_L = \frac{(\mu_{\text{Match}} - \mu_{\text{Distractor}})^2}{\sigma_{\text{noise}}^2} = \frac{(\mu_{\text{Match}} - \mu_{\text{Distractor}})^2}{\mu} \tag{9}$$

To fit a nonlinear model (**Fig. 6a**), we defined the nonlinearity, Φ, applied to each IT neuron as a monotonic piecewise linear function, with a threshold and saturation

$$\Phi(x_i) = \begin{cases} k_{\text{thr}} & \text{if } x_i < k_{\text{thr}} \\ x_i & \text{if } k_{\text{thr}} \le x_i \le k_{\text{sat}} \\ k_{\text{sat}} & \text{if } x_i > k_{\text{sat}} \end{cases} \tag{10}$$

where k_{thr} indicates the threshold value, k_{sat} indicates the saturation value and x_i indicates the mean response of the IT neuron to condition i. Note that if k_{thr} is lower than x_i and k_{sat} is larger than x_i for all conditions, then no nonlinearity is applied, so the formulation allows for the extreme case where $\Phi(x) = x$.

When applying this nonlinearity, we wished to avoid artificially creating information by applying transformations that could not be physically realized by neurons. Specifically, it is important to note that Linear-Nonlinear-Poisson (LNP) models operate by applying a nonlinearity to the mean neural responses across trials, and then simulate trial-by-trial variability with a Poisson process. In contrast, actual neurons can only operate on their inputs on individual trials, and their computations are therefore influenced by the trial-by-trial variability of their inputs. As an example, consider a toy neuron receiving only one input. When condition A is presented on three different trials, the neuron receives 7, 8, and 9 spikes. When condition B is presented on three trials, the neurons receives 8, 9 and 10 spikes. The mean input is thus 8 spikes for condition A and 9 spikes for condition B. An LNP model might attempt to take these inputs and apply a threshold at 8.5 spikes, below which it might set the firing rate to 0 spikes; such a nonlinearity would set the mean response to 0 spikes for condition A and 9 spikes for condition B, and after Poisson noise was regenerated, the distribution of responses for conditions A and B would be highly non-overlapping (for example, Poisson draws for condition A might be 0, 0 and 0 and Poisson draws for condition B might be 8, 9 and 10). However, artificially separating the input distributions in this way by a threshold violates laws of information processing. This can be demonstrated by noting that if the same threshold were applied trial by trial, it would produce 0, 0 and 9 spikes for condition A (mean = 3) and 0, 9 and 10 spikes for condition B (mean = 6.3), thereby preserving the fact that the two distributions are in fact overlapping. In our model, we aimed at exploiting the simplicity and expressive power of LNP models while also taking trial-by-trial response variability into consideration such that we did not artificially create information. Our strategy was twofold. First, we constrained the model by imposing that nonlinearities could only reduce the difference between the means of any pair of conditions. This was accomplished by imposing that matrix values could only be 'squashed' toward the threshold and the saturation, that is, values below the threshold are set to the value of threshold, and values above the saturation are set to the saturation value (see equation (10)). Second, we renormalized the response matrix after applying the nonlinearity to ensure that the overall SNR was not artificially increased by the generation of Poisson variability. In particular, we made the conservative assumption that the trial-by-trial variability was not modified by the nonlinearity, and therefore was equal to the mean response across all conditions before the application of the nonlinearity μ_{before} (see equation (9)). If the overall mean response was shifted by the nonlinearity to a new value μ_{after}, it was necessary to rescale the matrix to insure that the signal to noise ratio was consistent with the true variability, equal to μ_{before} (that is, no information was artificially created). This was accomplished by multiplying the response matrix by the ratio of μ_{after} and μ_{before}

$$\mathbf{M}_{\text{normalized}} = \mathbf{M} \cdot \frac{\mu_{\text{after}}}{\mu_{\text{before}}} \tag{11}$$

where $\mathbf{M}$ indicates the response matrix before normalization, and $\mathbf{M}_{\text{normalized}}$ is the response matrix after normalization.

When fitting the nonlinear model to our data (**Fig. 6a**), we explored all possible nonlinearities by allowing k_{thr} and k_{sat} to take any of the values in the original response matrix, for a total of 120 possible nonlinearities. The selected values were those that maximized the linearly separable target information (I_L, equation (9)).

Pairwise linear-nonlinear model of the transformation from IT to PRH. We created pairs of model PRH neurons via two orthonormal linear combinations of pairs of IT neurons, each followed by a static monotonic nonlinearity, that maximized the joint linearly separable information of the two model PRH neurons. Here we defined the response matrices of the two input IT cells as $\mathbf{I}_1$ and $\mathbf{I}_2$, the response matrices of the two output neurons as $\mathbf{O}_1$ and $\mathbf{O}_2$, the weights of the two linear combinations (indexed by input neuron, output pair) as w_{11}, w_{21}, w_{12} and w_{22}, and the two monotonic nonlinearities as Φ_1 and Φ_2.

$$\mathbf{O}_1 = \Phi_1(w_{11} \cdot \mathbf{I}_1 + w_{12} \cdot \mathbf{I}_2); \quad \mathbf{O}_2 = \Phi_2(w_{21} \cdot \mathbf{I}_1 + w_{22} \cdot \mathbf{I}_2) \tag{12}$$

where orthogonality of the weights was imposed by

$$w_{11} \cdot w_{21} + w_{12} \cdot w_{22} = 0 \tag{13}$$

and each pair of weights was constrained to a unitary norm

$$w_{11}^2 + w_{12}^2 = 1; \quad w_{21}^2 + w_{22}^2 = 1 \tag{14}$$

Because the weights were orthogonal and each pair was constrained to be unit norm, we could define the weights as a rotation matrix

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{15}$$

where θ is the angle by which the two-dimensional response space is rotated around the origin by the linear operation (**Fig. 7b**,**c**). Constraining the weights to be orthonormal is both necessary and sufficient to insure that no information is copied in the newly created neurons: the original space is simply rotated and the separation between the response clouds to different conditions are left intact. Conversely, non-orthogonal weights would result in 'copying over' the original information multiple times (note that copying the original information multiple times would not lead to an overall increase of the total information because the trial-by-trial variability in the two newly created neurons would be correlated). To find the optimal linear combinations for each pair of IT cells, we exhaustively explored all possible angles by systematically varying θ from 1–360 degrees. When responses were negative (that is, as a result of negative weights), we shifted the values of the response matrix to positive values and we renormalized the matrix to ensure that the shifting process did not artificially create information. This procedure resembles the renormalization we applied for static nonlinearities (equation (11)). First, we estimated the average trial-by-trial variability in the output matrix as the weighted combination of the average noise variances of the two input neurons

$$\sigma_O^2 = w_1^2 \cdot \sigma_{I1}^2 + w_2^2 \cdot \sigma_{I2}^2 \qquad (16)$$

where σ_O^2 is the noise variability in the output neuron, w_1 and w_2 are the weights, and σ_{I1}^2 and σ_{I2}^2 are the noise variances of the two input neurons. Next, we normalized the shifted response matrix $\mathbf{M}_{\text{shifted}}$ by multiplying it by the ratio between its mean response μ_{shifted}, and the actual predicted output noise σ_O^2

$$\mathbf{M}_{\text{normalized}} = \mathbf{M}_{\text{shifted}} \cdot \frac{\mu_{\text{shifted}}}{\sigma_O^2} \qquad (17)$$

This ensured that the overall SNR could not be influenced by changes in the mean response (that is, average noise variance under the Poisson assumption) as a result of the nonlinearity or the shift required to make all response values non-negative.

When considering our input population, we allowed for 'shifted copies' of our recorded IT neurons. More specifically, we allowed the model to make one selection from the set defined by each actual IT matrix we recorded and the 23 permutations of that matrix that are obtained by simultaneously shifting the four rows and four columns of the matrix. This procedure preserved the rules of combination between visual and working memory information (that is, the strengths of visual and cognitive modulation and their congruency; **Supplementary Fig. 4**), but shifted their object preferences. Stated differently, our assumption was that the rules of combination of visual and working memory signals were not specific to the object preferences of a neuron (that is, the brain does not employ one rule of combination for apple preferring neurons and a different rule for banana preferring neurons) and that any inhomogeneities with regard to object preferences that were included in our data set (for example, an excess of selective match detectors for object 1 as compared to object 4) were a result of finite sampling. For every possible pair of IT neurons, we generated all possible output neurons by considering all 24 matrix permutations, each paired by 360 possible angles, and each of those with all 120 possible nonlinearities. We also searched similar parameters for all possible pairs of output neurons generated by orthogonal weights to determine the pairing parameters that produced maximal joint linearly separable information.

Having determined the best parameters for every possible pair of IT neurons, we selected the subset of pairings that produced a model PRH population with the maximal amount of total linearly separable information while only allowing each IT input neuron to contribute to the model output population once. This selection problem can be reduced to an integer linear programming problem[44], and we implemented a standard solution using the GLPK library (http://www.gnu.org/software/glpk).

The role of asymmetric tuning correlations in untangling. Upon establishing that the pairwise linear-nonlinear model was effective at transforming nonlinearly separable information into a linearly separable format (**Fig. 6**), we were interested in an intuitive (and yet quantitatively accurate) understanding of how the model worked. Given any neuron's response matrix, one crucial property that enables a monotonic nonlinearity to extract linearly separable information (that is, to increase the distance between the mean response to the matches and the mean response to the distractors) is the degree to which the "tails" of the match and distractor distributions are non-overlapping (**Fig. 7a**). Although one could, in theory, fully characterize the match and distractor distributions and arrive to a closed-form estimate of the maximum extractable linearly separable information in a neuron's matrix via a nonlinearity, we focused on producing a simple estimate of this quantity based just on the first two moments of these distributions (that is, their means and variances). We postulated that the absolute value of the difference in variance across the matches (σ_{Match}^2) and the variance across the distribution of distractors ($\sigma_{\text{Distractor}}^2$) is a good predictor of the amount of linearly separable information that can be extracted by a monotonic nonlinearity (Δ_{info})

$$\Delta_{\text{info}} \approx k \cdot \left| \sigma_{\text{Match}}^2 - \sigma_{\text{Distractor}}^2 \right| \qquad (18)$$

where k is a proportionality constant. This estimate assumes that the means of the match and distractor distributions are the same and that variance differences thus translate into regions in which the high-variance distribution extends beyond the low-variance distribution (**Fig. 7a**). An improvement of this estimate could be obtained by correcting for the fact that the initial distance between the means of the two distributions (that is, the amount of pre-existing linearly separable information) always decreases the amount of overlap and thus always limits the amount of further information that can be extracted

$$\Delta_{\text{info}} \approx k \cdot \max\left(0, \Delta\sigma^2 - (\Delta\mu)^2\right) \qquad (19)$$

To extend the prediction to pairs of neurons, one must consider the covariance matrix for the bivariate distribution of match responses Σ_{Match} and of distractor responses $\Sigma_{\text{Distractor}}$, which can be further decomposed into the variances across matches and distractors and the tuning correlations for matches and distractors between the two neurons. Because the amount of linearly separable information gained by a pairing is proportional to the absolute value of the difference of the variances for matches and distractors ($\Delta\sigma^2$, equation (18)), the model will tend to pair IT neurons that maximize $\Delta\sigma^2$. Here we derived the amount of $\Delta\sigma^2$ that results from a pairing. First, we computed the variance across match responses $\sigma_{\text{Match, linear combination}}^2$ for a linear combination with weights w_1 and w_2 as

$$\sigma_{\text{Match, linear combination}}^2 = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \cdot \Sigma_{\text{Match}} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \cdot \begin{bmatrix} \sigma_{\text{Match},1}^2 & \rho_{\text{Match}} \cdot \sigma_{\text{Match},1} \cdot \sigma_{\text{Match},2} \\ \rho_{\text{Match}} \cdot \sigma_{\text{Match},1} \cdot \sigma_{\text{Match},2} & \sigma_{\text{Match},2}^2 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} =$$

$$= w_1^2 \cdot \sigma_{\text{Match},1}^2 + w_2^2 \cdot \sigma_{\text{Match},2}^2 + 2 \cdot w_1 \cdot w_2 \cdot \rho_{\text{Match}} \cdot \sigma_{\text{Match},1} \cdot \sigma_{\text{Match},2} \qquad (20)$$

Analogously, we computed the variance across distractors for the linear combination $\sigma_{\text{Distractor, linear combination}}^2$ as

$$\sigma_{\text{Distractor, linear combination}}^2 = w_1^2 \cdot \sigma_{\text{Distractor},1}^2 + w_2^2 \cdot \sigma_{\text{Distractor},2}^2 \qquad (21)$$
$$+ 2 \cdot w_1 \cdot w_2 \cdot \rho_{\text{Distractor}} \cdot \sigma_{\text{Distractor},1} \cdot \sigma_{\text{Distractor},2}$$

Consequently, we obtained the difference between variances by subtracting equation (21) from (20)

$$\Delta\sigma_{\text{linear combination}}^2 = w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 + 2 \cdot w_1 \cdot w_2 \cdot$$
$$\left(\rho_{\text{Match}} \cdot \bar{\sigma}_{\text{Match}}^2 - \rho_{\text{Distractor}} \cdot \bar{\sigma}_{\text{Distractor}}^2 \right) \qquad (22)$$

where $\Delta\sigma_1^2$ indicates the match/distractor variance difference for input neuron 1, $\Delta\sigma_2^2$ indicates the variance difference for input neuron 2, $\bar{\sigma}_{\text{Match}}^2$ is the geometric mean of the variances for matches of the two neurons, and $\bar{\sigma}_{\text{Distractor}}^2$ is the geometric mean of the variances for distractors. It is evident from equation (22) that variance difference between matches and distractors after pairing can derive from two different sources. First, variance differences can be inherited from the input neurons ($\Delta\sigma_1^2$ and $\Delta\sigma_2^2$)

$$\Delta\sigma_{\text{linear combination}}^2 \approx w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 \qquad (23)$$

For this type of variance difference, pairing is not required as linearly separable information could be extracted by applying a nonlinearity to each of the input matrices individually (**Fig. 7a**). Second, variance differences that did not exist in the inputs can be produced via asymmetric tuning correlations for matches and distractors

$$\Delta\sigma_{\text{linear combination}}^2 \approx 2 \cdot w_1 \cdot w_2 \cdot \left(\rho_{\text{Match}} \cdot \bar{\sigma}_{\text{Match}}^2 - \rho_{\text{Distractor}} \cdot \bar{\sigma}_{\text{Distractor}}^2 \right) \qquad (24)$$

As demonstrated in **Figure 6a**, the ability of the pairwise linear-nonlinear model to extract linearly separable information relied heavily on this second source of variance difference (compare the nonlinear model to the linear-nonlinear model). Finally, a prediction of how these variance differences translate into increases in linearly separable information could be made by applying equation (19) with the empirically derived constant of $k = 0.15$ applied to all pairs. Despite the great simplicity of this description and the fact that only the first two moments (mean, variance and covariance) of the match and distractor distributions are considered, this estimate was quite reliable at predicting the gain in linearly separable information in the model (Pearson correlation between the increase in linearly separable information for each linear-nonlinear model pair and the prediction (equation (19), $r = 0.84$, $r^2 = 0.7$)).

Statistical tests. For each of our single neuron measures, we reported P values as an evaluation of the probability that differences in the mean values that we observed in IT versus PRH were due to chance. As many of these measures were not normally distributed, we calculated these P values via a bootstrap procedure[45]. On each iteration of the bootstrap, we randomly sampled the true values from each population, with replacement, and we computed the difference between the means of the two newly created populations. We computed the P value as the fraction of 1,000 iterations on which the difference was flipped in sign relative to the actual difference between the means of the full data set (for example, if the mean for PRH was larger than the mean for IT, the fraction of bootstrap iterations in which the IT mean was larger than the PRH mean).

41. Wang, P. & Nikolic, D. An LCD monitor with sufficiently precise timing for research in vision. *Front. Hum. Neurosci.* **5**, 85 (2011).
42. Kelly, R.C. *et al.* Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. *J. Neurosci.* **27**, 261–264 (2007).
43. Averbeck, B.B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95**, 3633–3644 (2006).
44. Edmonds, J. & Johnson, E.L. Matching: a well-solved class of integer linear programs. in *Combinatorial Structures and Their Applications: Proceedings* (ed. Guy, R.K.) (Gordon and Breach, Calgary, 1970).
45. Efron, B. & Tibshirani, R.J. *An Introduction to the Boostrap* (CRC Press, 1994).

doi:10.1038/nn.3433