Published in final edited form as:
Nat Neurosci. 2013 August ; 16(8): 1132–1139. doi:10.1038/nn.3433.

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

Marino Pagan, **Luke S. Urban**, **Margot P. Wohl**, and **Nicole C. Rust**
Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

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 (IT) and perirhinal (PRH) cortex as macaque monkeys performed a task that required them to find targets within 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 (i.e. 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 “untangling” computations in PRH reformat input from IT by combining neurons with asymmetric tuning correlations for target matches and distractors.

Introduction

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 little-understood. Working memory signals are thought to be maintained in higher-order structures, such as prefrontal cortex (PFC), and

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.

these signals are thought to be “fed back” to earlier structures for combination with visual information [e.g. 3, 5, 6] although see [4]. The initial combination of visual and working memory signals is likely to occur within higher stages of the ventral visual pathway (e.g. V4 and inferotemporal cortex, 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 as well as the identity of a sought target [7–14]. While many models incorporate the simplifying assumption that the initial combination is implemented similarly by all neurons (e.g. 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 as 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 one that results in a non-linearly 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 demonstrated via lesioning studies to play a fundamental role in visual target search tasks [18, but see 19]. Our results demonstrate 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 to 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 (e.g. 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 as distractors in different blocks of trials (Fig. 2a). Behavioral performance was high overall (monkey 1: 94% correct; monkey 2: 92% correct; see Supp. Fig. 1a for performance as a function of trial position). 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 were generally looking for specific images as opposed to detecting the repeated presentation of any image [consistent with 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, unless otherwise stated, we counted spikes after the onset of each test (i.e. non-cue) stimulus within a window that accounted for neural latency but also preceded the monkeys’ reaction times (80 – 270 ms; see Methods and Supp. Fig. 1b for reaction time distributions). We then screened for neurons that were significantly modulated across the 16 conditions, as assessed by a one-way ANOVA (see Methods). Unless otherwise stated, our analyses were based on the data from correct trials.

We note that the three components of this task (described above) each 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 translates to diagonal structure (Fig. 2b, “four-object target detector”, “single-object target detector”, and “suppressed four-object target detector”). We find the “four-object target detectors” particularly compelling, as their matrix structure reflects the solution to the monkeys’ task (i.e. these neurons fire differentially when an image is viewed as a target versus as a distractor, and they do so for all four images included in the experiment; see also Fig. 4c). We also note that these examples of relatively pure selectivity existed within IT and PRH populations that were largely heterogeneous mixtures of different types of information (e.g. Fig. 2b, the “distractor detector”, which fires when image 2 is the stimulus and image 3 is the target, and the “mixture” neuron).

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 (i.e. the solution to the monkey’s task)? To assess this relationship, we began by probing the amount of “untangled” target match/distractor information in the IT and PRH populations with a linear read-out (Fig. 3a, right). 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 (SVM) and we then tested the boundary with separately measured trials (see Methods; Equation 1). Cross-validated population performance was significantly higher in PRH than in IT (Fig. 3b, left) and this result was confirmed in each monkey individually (Supp. 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” (Supp. Fig. 2a, “Adaptation control”) nor by changes in reward expectation as a function of the number of distractors encountered thus far in a trial or other position effects (Supp. Fig. 2a, “Position control”). Finally, while the analyses described thus far assume trial-by-trial independence between neurons, correlated variability has been shown to impact 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 (Supp. 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; see Supp. Fig. 2c for a similar analysis but based on trials grouped by reaction time). We also found that linear classification performance on error trials as compared to correct trials trended toward lower values in IT and was significantly lower in PRH (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 to the IT population? We computed a single-neuron measure of linearly separable target match information ("I_L") as a function of the separation of the responses to the target match and distractor conditions (Fig. 4c, inset; see Methods, Equation 3). We note that this measure maps directly onto the amount of "diagonal structure" in a neuron's response matrix (see Methods) and thus an idealized "four-object target detector" will 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 presented in Fig. 3b (left), 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 N best neurons. In PRH, the best neurons were indeed "four-object target detectors" (Fig. 4c, right) and performance saturated fairly quickly as a function of N (Supp. Fig. 2d, left). In contrast, in IT we found that the best neurons were detectors for at most two objects as targets (Fig. 4c, right) and IT performance was lower than PRH performance for equal-sized N (Supp. Fig. 2d, left). These results suggest that the compelling "four-object target detectors" we found in PRH were responsible for a large portion of the population performance differences we uncovered between IT and PRH. However, even after removing the best N neurons (as many as 23) from PRH, performance in PRH remained higher than IT (Supp. Fig. 2d, right). Notably, many of the top 23 PRH neurons had single-object target detector structure (Fig. 4c, right). 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 as well as 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 readout could reflect more total task-relevant information in PRH (i.e. 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 as compared to PRH (e.g. Fig. 3a, center versus right). 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 depends only on the degree to which the response clouds corresponding to target match and distractor conditions are non-overlapping, but not on the specific manner in which the response clouds are positioned relative to one another (Fig. 3a, compare center and right). As a measure of the total information available for match/distractor discrimination in the IT and PRH populations, we performed a cross-validated, ideal observer match/distractor classification of the population response on individual trials (see Methods, Equation 2).

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

Evidence for feed-forward “untangling” between IT and PRH

More “untangled” target match information in PRH as compared to 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-dependant 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. Importantly, these schemes can be distinguished via their predictions about the relative amounts and/or the timing of cognitive information in IT as compared to 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 to 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 as compared to 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 as compared to 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 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: 1) changing the visual image, 2) changing the cognitive context, and 3) noise due to trial-by-trial variability (see Methods, Equation 4). We found that cognitive modulations were approximately equal in strength in IT and PRH and that these modulations arrived slightly earlier in IT as compared to 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 also below). A decomposition of the combined cognitive signal into its linear ("working memory") and nonlinear (i.e. interaction) components revealed that, consistent with other reports [e.g. 22], working memory signals during the delay period ("persistent activity") are present but are weak in both areas (Supplementary Fig. 3c). Additionally, the nonlinear component predominated during the stimulus-evoked response period (Supplementary Fig. 3c), consistent with either working memory signals that combine nonlinearly (e.g. multiplicatively) with visual signals in these areas [e.g. 23] or with visual and working memory combinations that are inherited from elsewhere (e.g. V4). We describe these nonlinear signals in more detail in the next section.

We do acknowledge that the results we present here 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 (i.e. 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). Additionally, similar to other hierarchical descriptions of information processing [e.g. 16, 24, 25], we do not know that PRH receives its information via a direct projection from IT to PRH (e.g. 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 Fig. 5c can quantitatively account for our recorded responses. Similar to other functional model descriptions [e.g. 23–25, 26, 27, 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 (i.e. from IT to PRH), even in the absence of certainty regarding its exact biological implementation [29].

Taken together, the results reported in Figures 3–5 are consistent with a functional model in which visual and working memory signals are initially combined within 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 (i.e. from V1 to V4 to IT) for invariant object recognition [16, 30–32], and thus suggest that the brain transforms information into a manner that can be accessed via a linear population read-out not only for perception (i.e. identifying the content of a currently-viewed scene), but also for more cognitive tasks (i.e. finding a specific sought target object).

A pairwise LN model can account for PRH untangling

Next we were interested in evaluating whether an "untangling" transformation from IT to PRH could provide an accurate quantitative account of our data. We thus 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 began by ruling out *a priori* the class of models in which IT neurons combine linearly to produce PRH cells because we know that linear operations can move linearly separable information around within a population (i.e. between neurons) but cannot transform non-linearly separable information into a linearly separable format. Thus we began by testing the class of nonlinear models in which a static nonlinearity (i.e. thresholding and saturation; see Methods, Equation 10) 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, right, "N Model").

Next we considered the class of models in which pairs of IT neurons combine via a linear-nonlinear model ("LN model") to produce the responses of pairs of PRH cells (Fig. 6b). In fitting our model, we imposed the important constraint that information could not be replicated multiple times in the transformation from IT to PRH (i.e. 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 the pair of IT inputs (e.g. $(+\sqrt{0.5},+\sqrt{0.5})$ and $(+\sqrt{0.5},-\sqrt{0.5})$) and each IT neuron was included only once (see Methods, Equations 12–14). We searched all possible pairwise combinations of IT neurons and nonlinearities and selected the combinations that produced the largest gains in linearly separable information (see Methods). The resulting LN 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/distractor task with an ideal observer read-out (Fig. 6a, "LN Model"). The LN 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 (i.e. alignment) of visual and target signals (Supp. 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 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 lended itself to an exploration of the specific computational mechanisms underlying untangling, as described below.

Untangling relies on asymmetric tuning correlations in IT

To understand how the pairwise LN 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. As described in Fig. 7a, a nonlinearity can be effective in situations when the variance (i.e. the "spread") across one set of conditions (e.g. the matches; Fig. 7a, red solid) is higher than the other set (e.g. the distractors; Fig. 7a, gray). In such scenarios, the nonlinearity can change a subset of responses within the high variance set and thus increase the difference between the mean response to matches and distractors (Fig. 7a, red dashed vs gray); this translates into an increase of the amount of linearly separable target match information (see Methods Equation 19 for a more extensive description of the conditions required). Our results (Fig. 6a) suggest that pairing plays an important role in producing linearly separable information (as compared to 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 Fig. 3a (i.e. 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 hence 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 (i.e. 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 (e.g. 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, and thus 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 have formalized the intuitions presented in Fig. 7 into a quantitative prediction of the amount of linearly separable information that can be gained by pairing any two IT neurons via an LN 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 (see Methods, Equations 22, 24). Empirically we found that this prediction provided a good account of the linearly separable information extracted by our LN 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 LN model "untangled" 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 (e.g. matches) and a perfect tuning anti-correlation for the other set (e.g. distractors). However, we note that modest tuning correlation asymmetries are also predicted to translate into increases in linearly separable information (under appropriate conditions; see Methods, Equation 19). We found that our model did largely rely on modest (as opposed to maximal) tuning correlation asymmetries (Supp. 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 (Supp. 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). While 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: 1) discriminate between models that describe where and how visual and target signals combine (Fig. 5), 2) 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. 3,4,6, Supp. Fig. 4); and 3) provide a neural mechanism that can account for the untangling or reformatting process (Fig. 7, Supp. Fig. 5). Notably, our results are not predictable from earlier reports. Specifically, a series of groundbreaking studies reported signals that differentiate target matches from distractors not only in PRH [15, 33], but also in V4 [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 (e.g. V4 and IT) as compared to working memory inputs directly to PRH. While we can not definitively rule out the latter hypothesis, our results demonstrate 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, here we provide both a computational (i.e. “untangling”) and mechanistic (i.e. “pairing via asymmetric tuning correlations”) description of how that information might be reformatted within a feed-forward scheme.

While 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 demonstrate that nearly all the information for this task found in PRH is 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 are consistent with this description (Fig. 5e). Finally, our results demonstrate that a simple linear-nonlinear model can account for the transformation (Fig. 6a, Supp. 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 (i.e. IT) to a higher brain area (i.e. PRH). Similar to other functional model descriptions [e.g. 23–25, 26, 27, 28], we cannot rule out alternate proposals that predict the same neural responses but have different pathways (e.g. 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 (e.g. to produce the distractor detector in Fig. 2b; see also Supp. Fig. 4, right column). Similar incongruent neurons have also been reported in other studies [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 (i.e. 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 (e.g. "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 within PRH by combining IT neurons with asymmetric tuning correlations. Similar to other functional models [e.g. 23, 24, 26, 27, 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? While simple pairwise combinations of IT neurons were sufficient to explain the responses 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 [e.g. 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 (i.e. a linear) read-out in PRH as compared to IT. While 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 (e.g. 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. While 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 (e.g. from PRH, signals might propagate more deeply into the temporal lobe) and might be used for different purposes.

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 animal. 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 (e.g. the putamen and amygdala) to reach PRH. The region of IT recorded was located on both the ventral superior temporal sulcus (STS) and the ventral surface of the brain, over a 4 mm medial-lateral region located lateral to the anterior middle temporal sulcus (AMTS) that spanned 14–17 mm anterior to the ear canals [12, 30]. The region of PRH recorded was located medial to the AMTS 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, Inc.) and 16- and 24-channel U-probes with recording sites arranged linearly and separated by 150 micron spacing (Plexon Inc.). Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored via the OmniPlex Data Acquisition System (Plexon, Inc.). We performed all spike sorting manually offline using commercially available software (Plexon, Inc.). While 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. Additionally, our offline spike sorting procedures were performed blind to the specific experimental conditions (i.e. 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 [e.g. 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 4 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 within short blocks of ~1.7 minutes that included an average of 9 correct trials. Within each block, 4 presentations of each condition (for a fixed target) were collected and all four target blocks were presented within a “metablock” in pseudorandom order before reshuffling. A minimum of 5 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 ms to 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 reported here (i.e. claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys, see also 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 two times the standard deviation 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 (e.g. due to lower firing rates in one of our populations). The most liberal screening procedure 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 Fig. 4b and Supplementary Fig. 2b, these are the criteria we used for the Results. 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 (i.e. claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys).

To assess the impact of simultaneous trial-by-trial variability (i.e. "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 to 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 a population “response vector” **x** with a dimensionality equivalent to Nx1. 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 (4 trials) to test the representation (“test trials”). We tested two types of classifiers:

Linear classification - SVM—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 by [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, left). The linear readout took the following form:

$$f(x)=w^T x+b \quad (1)$$

where **w** is a Nx1 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 (SVM) 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:

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

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, right).

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 4 target match conditions and 12 distractor conditions; on each iteration we randomly selected 1 distractor condition from each image (for a total of 4 distractor conditions) to avoid artificial overestimations of classifier performance that could be produced by taking the prior distribution into account (e.g. 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 standard deviation, respectively, across 200 resampling iterations.

To assess the impact 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 standard deviation 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 standard deviation 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 measures of task-relevant information

Single-neuron measure of linearly separable target match information—As a single-neuron measure of match/distractor linear discriminability, we computed how well a neuron could linearly separate the responses to 4 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 $\mu_{Distractor,}$ divided by the variance of the spike count across trials, averaged across all 16 conditions σ^2_{noise} [43]:

$$I_L=\frac{(\mu_{Match}-\mu_{Distractor})^2}{\sigma^2_{noise}} \quad (3)$$

Single-neuron measures of visual and cognitive information

We began by performing a two-way analysis of variance (ANOVA), to parse each neuron's total response variability σ^2_{tot} (i.e. 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}:

$$\sigma^2_{tot}\cdot\nu_{tot}=\sigma^2_{vis}\cdot\nu_{vis}+\sigma^2_{\mathrm{targ}}\cdot\nu_{\mathrm{targ}}+\sigma^2_{NL}\cdot\nu_{NL}+\sigma^2_{noise}\cdot\nu_{noise} \quad (4)$$

where ν_{tot} = 319 (total number of degrees of freedom), ν_{vis} = 3 (degrees of freedom of visual modulation), ν_{targ} = 3 (degrees of freedom of target modulation), ν_{NL} = 9 (degrees of freedom of visual/target modulation interactions), ν_{noise} = 304 (degrees of freedom of noise variability). We then computed the ratios of signal modulations and noise variability to establish the 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, Supp. Fig. 3, Supp. Fig. 4), normalized by the noise variability, as: $\frac{\sigma^2_{vis}}{\sigma^2_{noise}}$. The fraction of a neuron's variance that could be attributed to changes in the target (i.e. working memory signal; Supp. Fig. 3) was captured by the variance of linear target modulations, normalized by the noise variability: $\frac{\sigma^2_{targ}}{\sigma^2_{noise}}$. The fraction of a neuron's variance that could be attributed to nonlinear cognitive modulation (Supp. Fig. 3) was captured by the variance of nonlinear interactions of visual and target identity, normalized by the noise variability: $\frac{\sigma^2_{NL}}{\sigma^2_{noise}}$. The fraction of a neuron's variance that could be attributed to overall changes in the cognitive context (i.e. overall cognitive signal; Fig. 5d-e, Supp. Fig. 4) was captured by the combined variance that could be attributed to linear and nonlinear target modulations, normalized by the noise variability: $\frac{\sigma^2_{\mathrm{targ}}+\sigma^2_{NL}}{\sigma^2_{noise}}$.

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, thus 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: $bias\approx\frac{\nu_{signal}}{n}$. 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" (i.e. 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 (i.e. the average response to each image as the visual stimulus) and the target marginal tuning (i.e. the average response to each image as the target):

$$\begin{gathered} lin\ congr = |\rho(x_{vis}, x_{\text{targ}})| \\ x_{vis}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R(vis{=}i, \text{targ}{=}k) \quad x_{\text{targ}}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R(vis{=}k, \text{targ}{=}i) \end{gathered} \quad (5)$$

where $R(vis = i$, targ $= k)$ 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 (i.e. congruent nonlinear combinations of visual and target signals) versus off the diagonal (i.e. 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 $\sigma^2_{nondiag}$:

$$\begin{gathered} \sigma^2_{diag} = (\mu_{Match} - \mu_{Distractor})^2/3 \\ \sigma^2_{nondiag} \cdot \nu_{nondaig} = \sigma^2_{NL} \cdot \nu_{NL} - \sigma^2_{diag} \cdot \nu_{diag} \end{gathered} \quad (6)$$

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

$$NL\ congr. = \frac{\sigma^2_{diag}}{\sigma^2_{diag} + \sigma^2_{nondiag}} \quad (7)$$

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:

$$I_{congr} = \frac{\sigma^2_{lin} \cdot lin\ congr + \sigma^2_{NL} \cdot NL\ congr}{\sigma^2_{lin} + \sigma^2_{NL}} \quad \text{where} \quad \sigma^2_{lin} = \sigma^2_{vis} + \sigma^2_{\text{targ}} \quad (8)$$

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 neuron’s 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).

Modeling the transformation from IT to PRH

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 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 thus compute I_L by replacing the noise variance term σ^2_{noise} with the mean responses across all conditions,μ:

$$I_L = \frac{(\mu_{Match} - \mu_{Distractor})^2}{\sigma^2_{noise}} = \frac{(\mu_{Match} - \mu_{Distractor})^2}{\mu} \quad (9)$$

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

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

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 thus their computations are 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 (e.g. Poisson draws for condition A might be 0, 0, 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), thus 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" towards the threshold and the saturation, i.e. 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 signal-to-noise ratio 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 $\mu_{after,}$ 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} (i.e. no information was artificially created). This was accomplished by multiplying the response matrix by the ratio of μ_{after} and $\mu_{before:}$

$$M_{normalized}=M\cdot\frac{\mu_{after}}{\mu_{before}} \quad (11)$$

where M indicates the response matrix before normalization, and $M_{normalized}$ is the response matrix after normalization.

When fitting the N 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 I_1 and $I_{2;}$ the response matrices of the two "output" neurons as O_1 and $O_{2;}$ 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.

$$O_1=\Phi_1(w_{11}\cdot I_1+w_{12}\cdot I_2)\ ;\ O_2=\Phi_2(w_{21}\cdot I_1+w_{22}\cdot I_2) \quad (12)$$

where orthogonality of the weights was imposed by:

$$w_{11}\cdot w_{21}+w_{12}\cdot w_{22}=0 \quad (13)$$

and each pair of weights was constrained to a unitary norm:

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

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

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

where θ is the angle by which the two-dimensional response space is rotated around the origin by the linear operation (compare Fig. 7b and 7c). 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 to 360 degrees. When responses were negative (i.e. 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:

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

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 $M_{shifted}$ by multiplying it by the ratio between its mean response $\mu_{shifted}$, and the actual predicted output noise σ_O^2:

$$M_{normalized}=M_{shifted}\cdot\frac{\mu_{shifted}}{\sigma_O^2} \quad (17)$$

This ensured that the overall signal-to-noise ratio could not be influenced by changes in the mean response (i.e. average noise variance under the Poisson assumption) due to 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 (i.e. the strengths of visual and cognitive modulation and their congruency; Supp. 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 (i.e. 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 (e.g. an excess of selective match detectors for object 1 as compared to object 4) were due to 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 LN 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 (i.e. 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 (i.e. their means and variances). We postulated that the absolute value of the difference in variance across the matches (σ^2_{Match}) and the variance across the distribution of distractors ($\sigma^2_{Distractor}$) is a good predictor of the amount of linearly separable information that can be extracted by a monotonic nonlinearity (Δ_{info}):

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

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 (i.e. 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:

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

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 $\Sigma_{\mathit{Distractor}}$, 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 $\sigma^2_{Match,lin.comb}$ for a linear combination with weights w_1 and w_2 as:

$$\begin{aligned} \sigma^2_{Match,lin.comb.} &= [w_1\, w_2] \cdot \sum\nolimits_{Match} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \quad \ldots \\ &= [w_1\, w_2] \cdot \begin{bmatrix} \sigma^2_{Match,1} & \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} \\ \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} & \sigma^2_{Match,2} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \quad \ldots \\ &= w_1^2 \cdot \sigma^2_{Match,1} + w_2^2 \cdot \sigma^2_{Match,2} + 2 \cdot w_1 \cdot w_2 \cdot \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} \end{aligned} \quad (20)$$

Analogously, we computed the variance across distractors for the linear combination $\sigma^2_{Distractor,lin.comb.}$ as:

$$\sigma^2_{Distractor,lin.comb.} = w_1^2 \cdot \sigma^2_{Distractor,1} + w_2^2 \cdot \sigma^2_{Distractor,2} + 2 \cdot w_1 \cdot w_2 \cdot \rho_{Distractor} \cdot \sigma_{Distractor,1} \cdot \sigma_{Distractor,2} \quad (21)$$

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

$$\Delta\sigma^2_{lin.comb.} = w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 + 2 \cdot w_1 \cdot w_2 \cdot \left(\rho_{Match} \cdot \overline{\sigma}^2_{Match} - \rho_{Distractor} \cdot \overline{\sigma}^2_{Distractor}\right) \quad (22)$$

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, $\overline{\sigma}^2_{Match}$ is the geometric mean of the variances for matches of the two neurons, and $\overline{\sigma}^2_{Distractor}$ 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 $\Delta\sigma_2^2$):

$$\Delta\sigma^2_{lin.comb.} \approx w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 \quad (23)$$

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:

$$\Delta\sigma^2_{lin.comb.} \approx 2 \cdot w_1 \cdot w_2 \cdot \left(\rho_{Match} \cdot \overline{\sigma}^2_{Match} - \rho_{Distractor} \cdot \overline{\sigma}^2_{Distractor}\right) \quad (24)$$

As demonstrated in Fig. 6a, the ability of the pairwise LN model to extract linearly separable information relied heavily on this second source of variance difference (compare the N model to the LN 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 LN 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 1000 iterations on which the difference was flipped in sign relative to the actual difference between the means of the full dataset (e.g. 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).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Eye Institute of the 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). We thank Yale Cohen, Tony Movshon, Eero Simoncelli and Alan Stocker for helpful discussions. We are especially grateful to David Brainard and Josh Gold for detailed comments on the work and on the manuscript. We also thank Jennifer Deutsch for technical assistance and Christin Veeder for veterinary support.

References

1. Salinas E. Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. J Neurosci. 2004; 24:1113–8. [PubMed: 14762129]
2. Salinas, E.; Bentley, NM. Gain modulation as a mechanism for switching reference frames, tasks, and targets. In: Josic, K.; Rubin, J.; Matias, M.; Romo, R., editors. Coherent behavior in neuronal networks. Springer; New York: 2009. p. 121-142.
3. Engel TA, Wang XJ. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. J Neurosci. 2011; 31:6982–96. [PubMed: 21562260]

4. Sugase-Miyamoto Y, Liu Z, Wiener MC, Optican LM, Richmond BJ. Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. PLoS Comput Biol. 2008; 4:e1000073. [PubMed: 18464917]

5. Miller EK, Erickson CA, Desimone R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. Journal of Neuroscience. 1996; 16:5154–5167. [PubMed: 8756444]

6. Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, Miyashita Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature. 1999; 401:699–703. [PubMed: 10537108]

7. Haenny PE, Maunsell JHR, Schiller PH. State Dependent Activity in Monkey Visual-Cortex.2. Retinal and Extraretinal Factors in V4. Experimental Brain Research. 1988; 69:245–259. [PubMed: 3345806]

8. Maunsell JHR, Sclar G, Nealey TA, Depriest DD. Extraretinal Representations in Area-V4 in the Macaque Monkey. Visual Neuroscience. 1991; 7:561–573. [PubMed: 1772806]

9. Bichot NP, Rossi AF, Desimone R. Parallel and serial neural mechanisms for visual search in macaque area V4. Science. 2005; 308:529–534. [PubMed: 15845848]

10. Chelazzi L, Miller EK, Duncan J, Desimone R. Responses of neurons in macaque area V4 during memory-guided visual search. Cereb Cortex. 2001; 11:761–72. [PubMed: 11459766]

11. Eskandar EN, Richmond BJ, Optican LM. Role of Inferior Temporal Neurons in Visual Memory.1. Temporal Encoding of Information About Visual Images, Recalled Images, and Behavioral Context. Journal of Neurophysiology. 1992; 68:1277–1295. [PubMed: 1432084]

12. Liu Z, Richmond BJ. Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules. Journal of Neurophysiology. 2000; 83:1677–1692. [PubMed: 10712488]

13. Gibson JR, Maunsell JHR. Sensory modality specificity of neural activity related to memory in visual cortex. Journal of Neurophysiology. 1997; 78:1263–1275. [PubMed: 9310418]

14. Lueschow A, Miller EK, Desimone R. Inferior Temporal Mechanisms for Invariant Object Recognition. Cerebral Cortex. 1994; 4:523–531. [PubMed: 7833653]

15. Miller EK, Desimone R. Parallel Neuronal Mechanisms for Short-Term-Memory. Science. 1994; 263:520–522. [PubMed: 8290960]

16. DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn Sci. 2007; 11:333–41. [PubMed: 17631409]

17. Suzuki WA, Amaral DG. Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. J Comp Neurol. 1994; 350:497–533. [PubMed: 7890828]

18. Meunier M, Bachevalier J, Mishkin M, Murray EA. Effects on Visual Recognition of Combined and Separate Ablations of the Entorhinal and Perirhinal Cortex in Rhesus-Monkeys. Journal of Neuroscience. 1993; 13:5418–5432. [PubMed: 8254384]

19. Buffalo EA, Ramus SJ, Squire LR, Zola SM. Perception and recognition memory in monkeys following lesions of area TE and perirhinal cortex. Learn Mem. 2000; 7:375–82. [PubMed: 11112796]

20. Cohen MR, Maunsell JH. Attention improves performance primarily by reducing interneuronal correlations. Nat Neurosci. 2009; 12:1594–600. [PubMed: 19915566]

21. Graf AB, Kohn A, Jazayeri M, Movshon JA. Decoding the activity of neuronal populations in macaque primary visual cortex. Nat Neurosci. 2011; 14:239–45. [PubMed: 21217762]

22. Fuster JM, Jervey JP. Inferotemporal Neurons Distinguish and Retain Behaviorally Relevant Features of Visual-Stimuli. Science. 1981; 212:952–955. [PubMed: 7233192]

23. Reynolds JH, Heeger DJ. The normalization model of attention. Neuron. 2009; 61:168–85. [PubMed: 19186161]

24. Rust NC, Mante V, Simoncelli EP, Movshon JA. How MT cells analyze the motion of visual patterns. Nat Neurosci. 2006; 9:1421–31. [PubMed: 17041595]

25. Gold JI, Shadlen MN. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron. 2002; 36:299–308. [PubMed: 12383783]

26. Simoncelli EP, Heeger DJ. A model of neuronal responses in visual area MT. Vision Res. 1998; 38:743–61. [PubMed: 9604103]

27. Heeger DJ. Normalization of cell responses in cat striate cortex. Vis Neurosci. 1992; 9:181–97. [PubMed: 1504027]

28. Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. J Opt Soc Am A. 1985; 2:284–299. [PubMed: 3973762]

29. Marr, D. Vision. Cambridge, MA: MIT Press; 1982.

30. Rust NC, DiCarlo JJ. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. J Neurosci. 2010; 30:12978–95. [PubMed: 20881116]

31. Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. Science. 2005; 310:863–6. [PubMed: 16272124]

32. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? Neuron. 2012; 73:415–34. [PubMed: 22325196]

33. Chelazzi L, Miller EK, Duncan J, Desimone R. A neural basis for visual search in inferior temporal cortex. Nature. 1993; 363:345–7. [PubMed: 8497317]

34. Maunsell JHR, Treue S. Feature-based attention in visual cortex. Trends in Neurosciences. 2006; 29:317–322. [PubMed: 16697058]

35. Rigotti M, Ben Dayan Rubin D, Wang XJ, Fusi S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. Front Comput Neurosci. 2010; 4:24. [PubMed: 21048899]

36. Najemnik J, Geisler WS. Optimal eye movement strategies in visual search. Nature. 2005; 434:387–91. [PubMed: 15772663]

37. Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol. 2001; 86:1916–36. [PubMed: 11600651]

38. Law CT, Gold JI. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nat Neurosci. 2009; 12:655–63. [PubMed: 19377473]

39. Lavenex P, Suzuki WA, Amaral DG. Perirhinal and parahippocampal cortices of the macaque monkey: projections to the neocortex. J Comp Neurol. 2002; 447:394–420. [PubMed: 11992524]

40. Minsky, M.; Papert, S. Perceptrons: An introduction to computational geometry. Cambridge, MA: MIT Press; 1969.

41. Wang P, Nikolic D. An LCD Monitor with Sufficiently Precise Timing for Research in Vision. Front Hum Neurosci. 2011; 5:85. [PubMed: 21887142]

42. Kelly RC, Smith MA, Samonds JM, Kohn A, Bonds AB, Movshon JA, Lee TS. Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. J Neurosci. 2007; 27:261–4. [PubMed: 17215384]

43. Averbeck BB, Lee D. Effects of noise correlations on information encoding and decoding. J Neurophysiol. 2006; 95:3633–44. [PubMed: 16554512]

44. Edmonds, J.; Johnson, EL. Matching: a well-solved class of integer linear programs. In: Guy, RK., editor. Combinatorial structures and their applications: proceedings. Gordon and Breach; Calgary: 1970.

45. Efron, B.; Tibshirani, RJ. An introduction to the boostrap. Boca Raton: CRC Press; 1994.

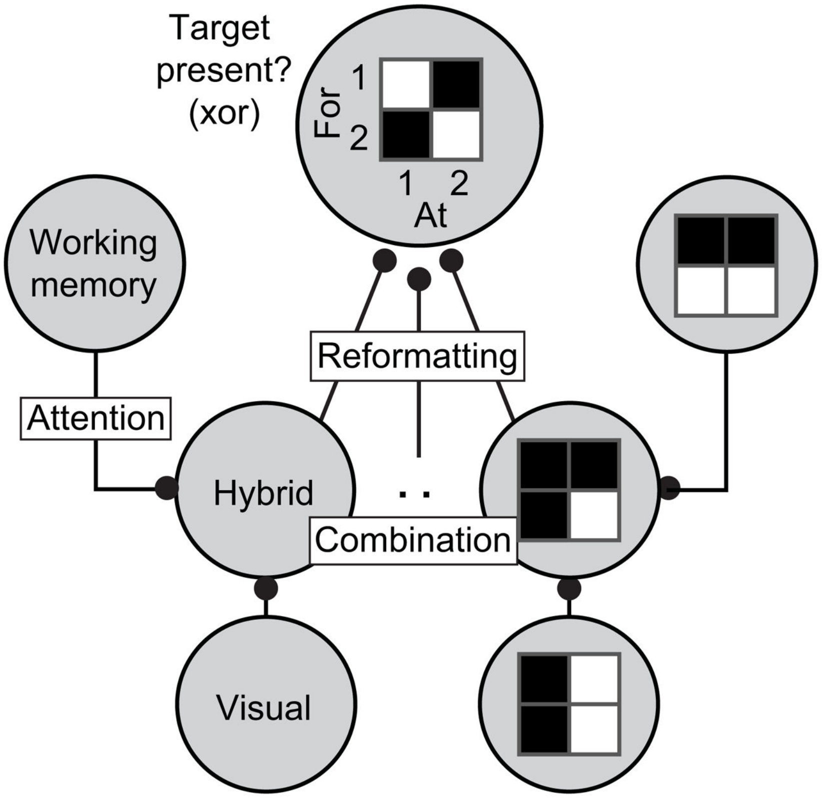

Figure 1. Theoretical proposals of the neural mechanisms involved in finding visual targets

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) nor (2,1). Producing such a signal requires at least two stages of processing in a feed-forward 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.

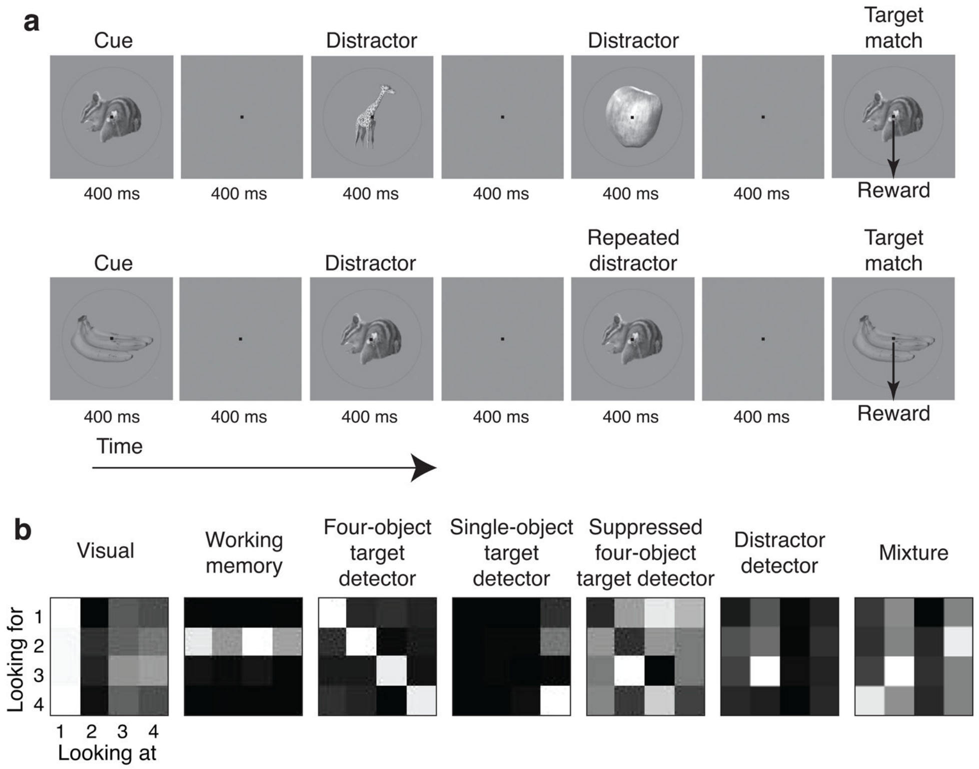

Figure 2. The delayed-match-to-sample (DMS) task and example neural responses

a) We trained monkeys to perform a DMS 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, IT. Single-neuron linearly separable information ("I_L"; Fig. 4c) values (left-to-right): 0.01, 0.02, 3.33, 0.39, 0.44, 0.01, 0.06.

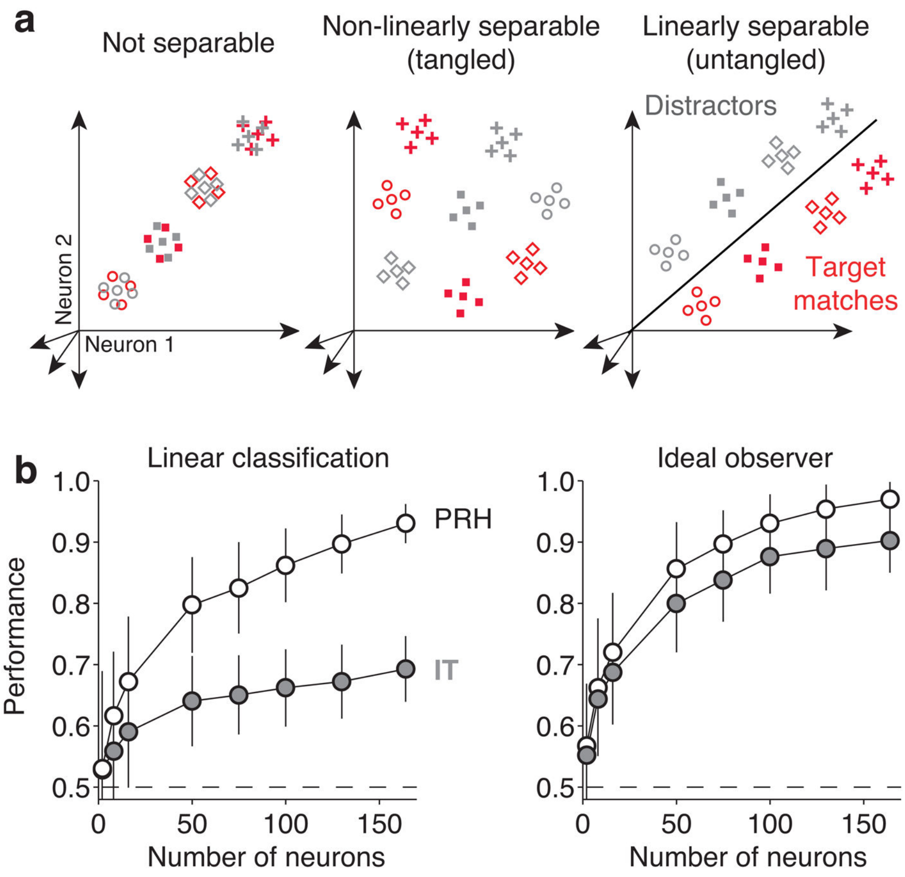

Figure 3. Population performance

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 due to trial-by-trial variability. The target-switching task (Figure 2) requires 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 (see Text, Methods). We measured total separability with a cross-validated, ideal observer analysis (see Text, Methods). 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.

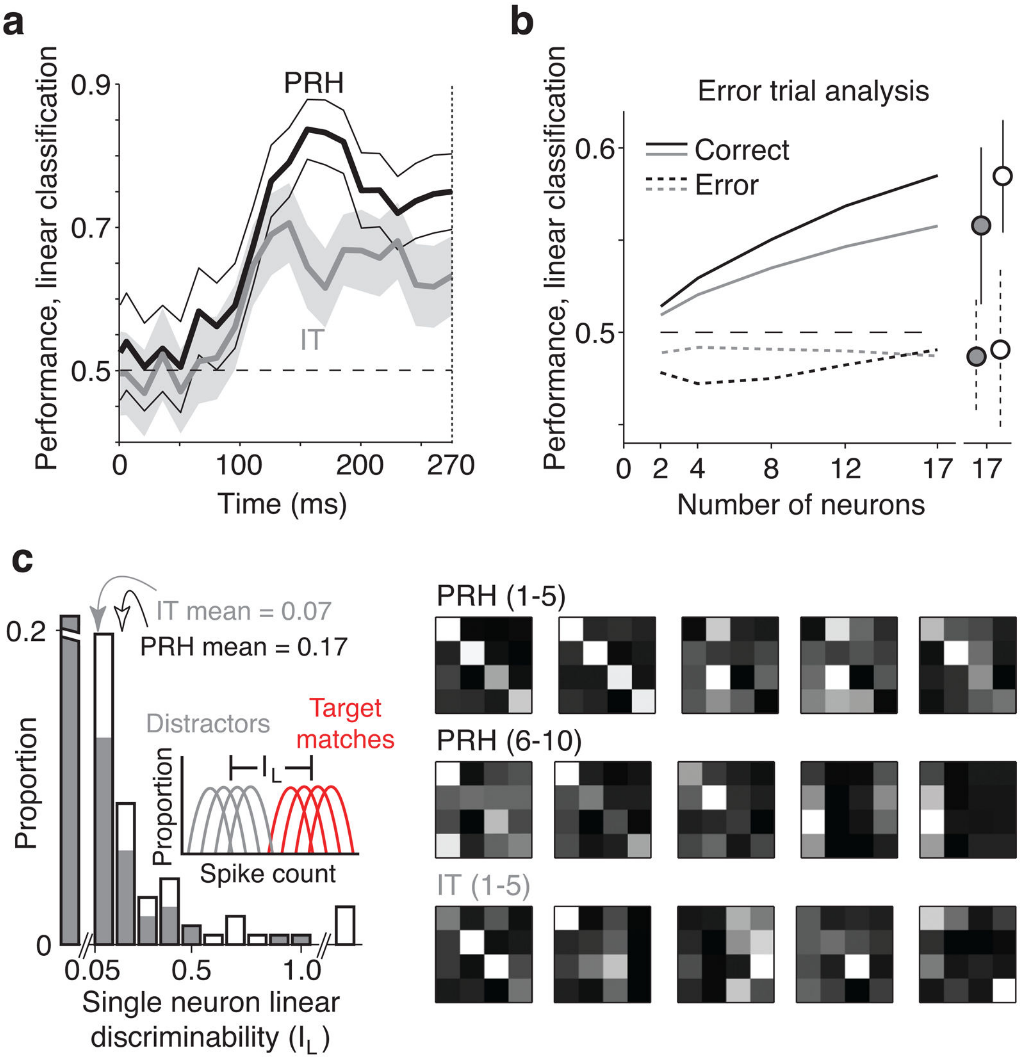

Figure 4. Additional population performance measures

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 dotted line indicates the minimum reaction time on these trials (270 ms). **b)** Linear classification performance on error (dotted) as compared to correct (solid) trials (same conventions as Fig. 3b, left; see 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 (i.e. responding to a distractor). We performed the analysis separately for each multi-channel recording session and then averaged across sessions. **c)** *Left*, Histograms of linearly separable target match information ("I_L"; see Methods Equation 3, computed for IT (gray) and PRH (white). Arrows indicate means. The last bin includes PRH neurons with I_L of 1.1, 1.4, and 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 shown with the same conventions as Fig. 2b and the rankings labeled.

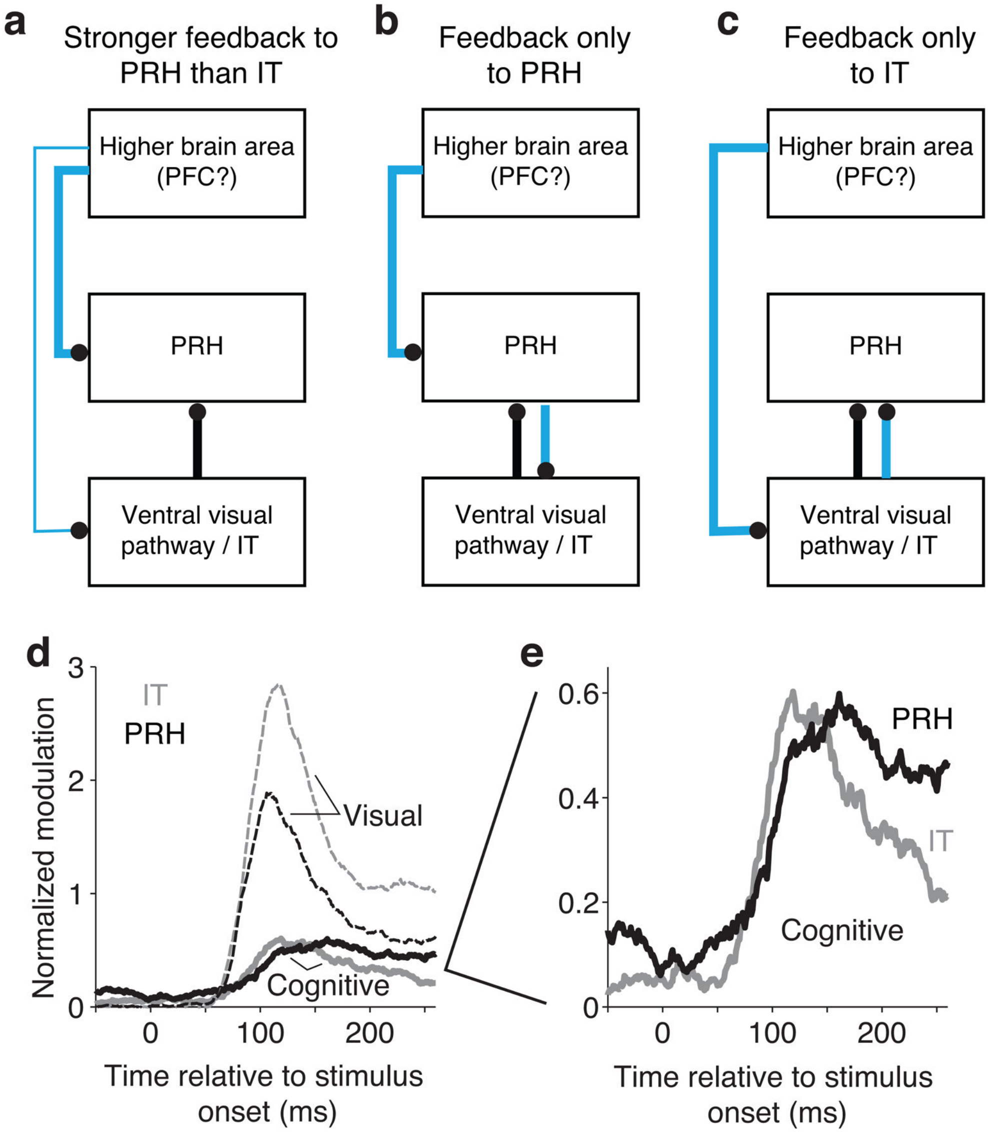

Figure 5. Discriminating between classes of models that predict more "untangled" target match information in PRH than IT

a–c) Black lines indicate visual input; cyan lines indicate "cognitive input" that can take the form of working memory or target match information (see Text). **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 (see Methods, Equation 4), 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; "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 subpanel d. In panels d and e, response matrices were calculated from spikes in 60 ms bins with 1 ms shifts between bins.

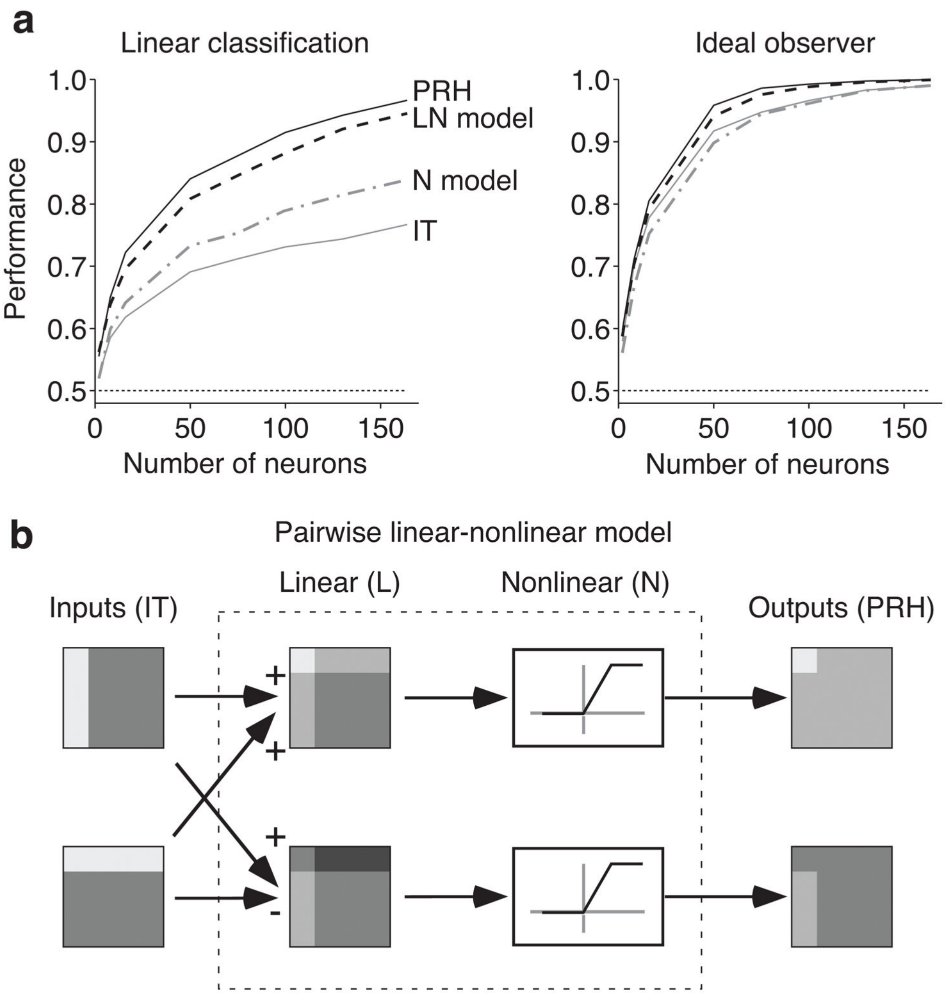

Figure 6. Modeling the transformation from IT to PRH

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), with the same conventions described 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 (LN model) we fit to describe the transformation from IT to PRH, shown for two idealized IT neurons. To create the LN model, pairs of IT neurons were combined via two sets of orthogonal linear weights, followed by a nonlinearity to create two model PRH neurons.

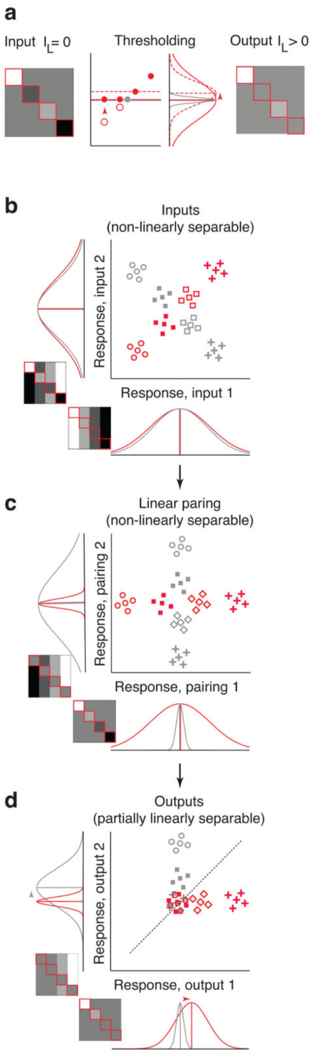

Figure 7. The neural mechanisms underlying untangling

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), thus 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 depicted in the same format as Fig. 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 within 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 (i.e. a difference) in the neurons’ tuning correlations for matches and distractors (Methods, Equation 24).