Current and future goals are represented in opposite patterns in object-selective cortex

Anouk M. van Loon[1,2], Johannes J. Fahrenfort[1,3] & Christian N. L. Olivers[1,2]

[1]*Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, The Netherlands*

[2]*Institute of Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*

[3]*Department of Brain and Cognition, University of Amsterdam, The Netherlands*

Address correspondence to:

Christian N. L. Olivers

Department of Experimental and Applied Psychology,

Vrije Universiteit Amsterdam,

Van der Boechorststraat 1,

1081 BT, Amsterdam,

The Netherlands.

Email: c.n.l.olivers@vu.nl

Acknowledgements:

We thank Assaf Harel for providing the pictures used as stimuli. This research was supported by the European Research Council (ERC) under grant agreement no. 615423 (to CNLO).

Conflict of Interest:

None declared.

Abstract

Adaptive behavior requires the separation of current from future behavioural goals in working memory. We used fMRI of object-selective cortex to determine the representational (dis)similarities of memory representations serving current and prospective perceptual tasks. Human participants remembered different target objects for two consecutive visual search tasks. A cue indicated whether an object should be looked for first (making it currently relevant), or second (making it prospectively relevant). Prior to the first search, category decoding for currently relevant targets was better than for prospectively relevant targets, while representations were similar. During the first search, the prospective target representation could also momentarily be decoded, but now current and prospective targets of the same category revealed anti-correlated voxel patterns. We propose that the brain separates current from prospective memories within the same neuronal ensembles through opposite representational patterns of activity versus activity-silent responsivity.

Introduction

Adaptive human behavior requires the representation of both imminent and future goals in response to changing task requirements. Little is known about how the brain distinguishes between information that is currently relevant and information that is only prospectively relevant.

While working memory is thought to be pivotal to the active maintenance of current task goals, representations serving prospective tasks should be shielded from affecting currently relevant input and output, and vice versa. Studies using reaction time and eye movement measures have indeed shown that currently and prospectively relevant representations in working memory differentially bias processing of perceptual input (e.g Carlisle & Woodman, 2011; Downing & Dodds, 2003; Houtkamp & Roelfsema, 2006; Mallett & Lewis-Peacock, 2018; Olivers & Eimer, 2011; van Loon, Olmos Solis, & Olivers, 2017). Furthermore, studies using multi-variate pattern analyses (MVPA) of functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data have shown that while memory items required for an upcoming memory test can be readily decoded, the evidence for items required for a prospective task temporarily drops to baseline levels until they become relevant again (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; LaRocque, Riggall, Emrich, & Postle, 2017; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). These and other findings have led to the hypothesis that items in working memory may adopt different states or representational formats (Barak & Tsodyks, 2014; D'Esposito & Postle, 2015; LaRocque, Lewis-Peacock, & Postle, 2014; Olivers, Peters, Houtkamp, & Roelfsema, 2011; Stokes, 2015). While currently relevant items are represented through patterns of firing across populations of neurons, prospectively relevant representations may

be stored in what has been referred to as an “activity-silent”, or “hidden” state. One way in which such a state can be achieved is through short-term potentiation of synaptic connectivity in the neuronal population, as induced by the initial firing activity during encoding and active storage within that same population (i.e. the short-term version of “what fires together wires together”; Erickson, Maramara, & Lisman, 2010; Mongillo, Barak, & Tsodyks, 2008). Another way is through changes in the membrane potentials of the previously firing neurons (e.g. Stokes, 2015). We will collectively refer to these options as changes in the *responsivity* (versus the *activity*) of a neuronal ensemble.

Such latent changes in responsivity are by definition difficult to test through activity-based measures. One prediction is that prospective memories may re-emerge in our dependent measures when *unrelated* activity is sent through the network and interacts with the pattern of changed responsivity that reflects the activity-silent memory. This is indeed what Rose and colleagues (2016) recently reported. They found that prospective memory representations which could initially no longer be decoded during a working memory delay period could successfully be reconstructed after applying a brief burst of transcranial magnetic stimulation (TMS, Rose et al., 2016). Likewise, Wolff and colleagues recently reported enhanced decoding of a memorized oriented grating shortly after observers were presented with a visual pattern that was neutral with respect to the memorized orientation (Wolff, Ding, Myers, & Stokes, 2015; Wolff, Jochim, Akyürek, & Stokes, 2017). However, although these studies show that there is information present on prospectively stored memories, it is as yet unclear what the representational format of such prospective memories is, and how this relates to currently relevant memories. A priori there appear to be a number of hypotheses.

First, the standard short-or long-term potentiation mechanism predicts that the altered pattern of responsivity directly follows the pattern of activity during encoding of the item, thus predicting a high degree of similarity between the active and the silent representation when revived. A second possibility is that it is unnecessary to assume activity-silent representations at all, as has recently been argued by (Schneegans & Bays, 2017). Instead, they argued for a single maintenance mechanism in which items in memory are stored through similar patterns of firing activity, with the only difference being the degree of activation. Their model simulations provide a proof of concept that the revival of a memory can be explained by selectively boosting the still present, but lowered activity, rather than by the reconstruction from hidden states of responsivity. Also under this scenario the same pattern of activation should emerge for current and prospective

memories, except for a difference in strength. The third possibility is that prospectively relevant items are stored in an altogether different pattern compared to actively maintained items – that is, they may be first transformed, stored in different neuronal populations, different layers or even different brain regions – whether through changed activity or responsivity. This was recently proposed by Christophel and colleagues (2018), who found currently relevant visual items to be more strongly represented in posterior brain areas (notably visual cortex), while prospectively relevant items were more strongly represented in frontal regions (notably the Frontal Eye Fields). Under this scenario the representational overlap between current and prospective items within the brain regions involved is expected to be minimal. Although crucial for current theories of working memory, so far, studies have not directly compared the representational pattern of current and prospective memories.

To further understand how working memory distinguishes between the now and the future, we investigated the similarity between currently relevant and prospectively relevant representations. Such a comparison requires the decoding of specific content representations within one and the same stimulus category and its associated brain area, when either currently or prospectively relevant. For this purpose, we used MVPA of fMRI activity in object-selective visual cortex. We asked observers to perform two consecutive visual searches for particular target objects drawn from different object categories (see Figure 1). Importantly, prior to the first search, a cue indicated whether the target object of interest would be relevant for the *first* search (turning it into a *current* target) or for the *second* search (turning it into a *prospective* target).

First, we replicate and extend the finding that currently relevant objects are represented more strongly than prospectively relevant objects by showing stronger category-selective responses for current targets than for prospective targets during the delay prior to search. Second, we find evidence for a momentary revival of the prospective search target while observers are performing the first search task, thus extending the demonstration that prospective memories can be reconstructed by sending unrelated activity through the network (Rose et al., 2016; Wolff, Jochim, Akyurek, & Stokes, 2017). Third, and most importantly, we find that when the prospective memory momentarily emerges during the first search, the pattern of activity directly relates to the pattern of activity evoked when it is currently relevant, but in an *inverse* fashion, thus making prospective patterns systematically *dis*similar from their current counterparts. These results suggest that prospective memories are protected from current memories by maintaining them in an opposite representational space.

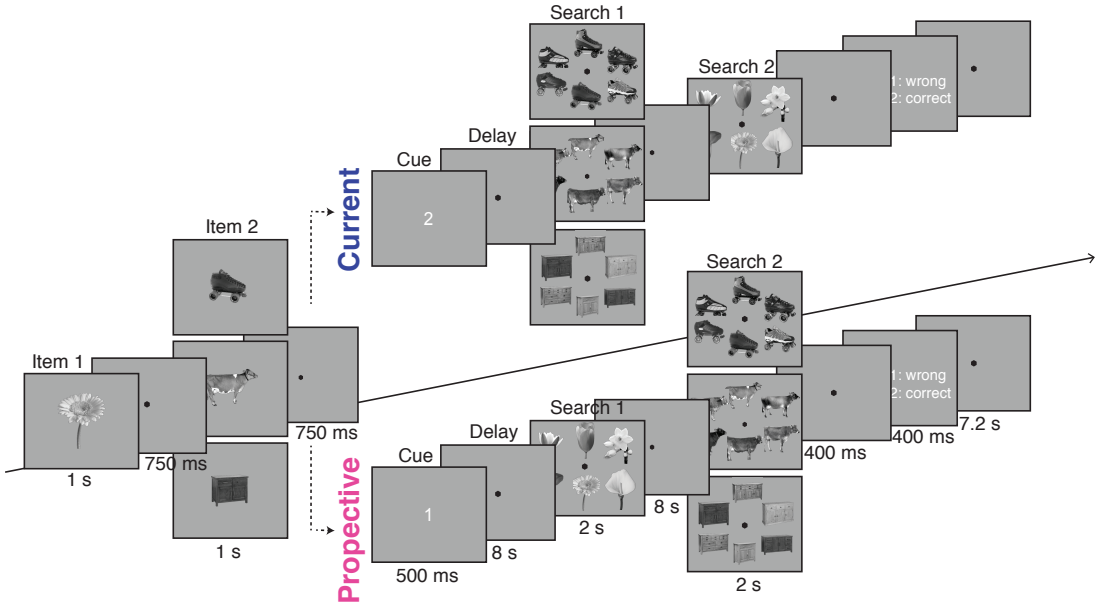


Figure 1. Trial design. *On each trial, participants performed two consecutive search tasks. The target objects for both of the search tasks were presented at the start of the trial. One of the objects could either be a cow, dresser or skate (with four exemplars for each category), and was used for the decoding analyses. The other target was always the same flower and was never part of the decoding scheme. The order of presentation (flower or category object) was counterbalanced across trials. A retro-cue ("1" or "2") determined whether the object category of interest was to be searched for first, rendering it currently relevant (the 'Current' item), or searched for second, rendering it prospectively relevant (the 'Prospective' item). The cue was followed by a delay, then the first search display, followed by a second delay and finally the second search display. During each search display, participants indicated whether the object was present or absent using a button press. Thus, in the Current condition, observer first searched for the object category of interest (cow, dresser, or skate), and then for the flower, while in the Prospective condition they first searched for the flower and then for the object category of interest. At the end of each trial and run participants received feedback about their performance.*

Results

To examine the relationship between currently and prospectively relevant representations, on each trial observers (N=24) performed two consecutive search tasks (search 1 and search 2), one for what we will refer to as the object category of interest (which was either a cow, a dresser, or a skate) and another one for a flower target. The two targets were presented at the start of each trial, after which a cue indicated whether the object category of interest (cow, dresser or skate) was to be searched for first, or second – thus making it currently or prospectively relevant. For each search, participants indicated whether the target object was present or absent among six exemplars of the same category. To limit the working memory load, and to maximize the chances of decoding the target category of interest (whether

current or prospective), the flower search task always involved searching for the same flower, while the target category of interest changed on every trial. Our primary analysis focused on posterior fusiform cortex (pFs), which is known to be involved in representing these object categories, and which we independently mapped for each participant (following Harel, Kravitz, & Baker, 2014; Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2013; Lee, Kravitz, & Baker, 2013; Malach et al., 1995).

Behavioral results

We computed mean RTs and mean accuracy for both types of search task when they came either first or second. Next, we ran a 2-way repeated measures ANOVA (N= 24) with factor Search (search 1 and search 2) and Object Category (flower and category targets (cow, dresser and skate), N = 24). As expected, search for the flower was overall faster and more accurate than for category targets, mean RT: 794 vs. 1387 ms (Search 1), and 772 vs. 1411 ms (Search 2) $F_{(1,23)} = 928.18$, $p < 0.001$; percentage correct 98.6 vs 82.2 (Search 1), 98.1 vs 76.0 (Search 2), $F_{(1,23)} = 183.06$, $p < 0.001$. Furthermore, the first search was more accurate than the second, $F_{(1,23)} = 10.87$, $p = 0.003$. There was also an interaction for both accuracy and speed ($F_{(1,23)} = 9.22$, $p = 0.006$ and $F_{(1,23)} = 6.32$, $p = 0.02$, respectively): the search for the category target had the lowest accuracy when performed second, while the flower search was fastest when second. Overall these results show that we were successful at minimizing the working memory load for the repeated (flower) target, thus maximizing the chances of decoding the primary target category.

fMRI results: Object category decoding as a function of current and prospective relevance

To investigate whether we could decode memory content for currently and prospectively relevant objects, we trained a classifier on the multivoxel response patterns in pFs using each target object category of interest (Cows, Dressers, Skates), for each TR. First, we trained and tested the classifier separately for trials in which the target category was relevant during Search 1 (currently relevant) and when the target category was relevant during Search 2 (prospectively relevant, see Methods section for details). Object category classification performance for this *within-relevance* decoding scheme is shown in Figure 2A. We focused our statistical analysis on the averaged classification performance for three intervals in the trial (of three TRs each), as predetermined on the basis of Lee and colleagues (2013), referred to as Delay, Search 1, and Search 2 (see Methods). We used paired t-tests (N = 24) to compare the classification performance from chance (33%) for these intervals, as

well as between Current and Prospective conditions. The average activity for these time windows is also shown in the top panel of Figure S3A.

While the within-relevance decoding scheme provides evidence for the presence of current and prospective representations, it does not reveal whether these representations are similar or different. Therefore, we additionally implemented a *cross-relevance* decoding scheme in which we trained the classifier when the objects were currently relevant and tested when the same objects were prospectively relevant (referred to as PC), and vice versa (referred to as CP, see Methods). Figure 2B shows the classification accuracy for this cross-relevance decoding schemes. Crucially, if current and prospective representations are similar, above-chance classification accuracy is expected. If representations are dissimilar in an unrelated fashion, classification is expected to be at chance levels, while below-chance classification is predicted when representations are dissimilar, but in a systematic, anti-correlated fashion.

The delay prior to the first search: Stronger decoding for current than for prospective targets, but similar representations.

As can bee seen in Figure 2A, during the Delay prior to search the within-relevance decoding resulted in significant above chance object category decoding both when currently relevant ($t_{(1,23)}$ = 8.18, p < 0.001, η_p^2 = 1.67, blue lines) and when prospectively relevant ($t_{(1,23)}$ = 5.67, p < 0.001, η_p^2 = 1.16, pink lines). Notably, decoding performance was higher when the item was currently relevant than when it was prospectively relevant (Current vs. Prospective: $t_{(1,23)}$ = 3.22, p = 0.004, η_p^2 = 0.66), consistent with its importance for the upcoming search task. Thus, object-selective cortex proves sensitive to object category as well as task-relevance prior to search.

Next, we used the *cross-relevance* decoding scheme to assess whether current and prospective targets shared the same neural representational pattern (see Figure 2B). This analysis revealed strong above-chance classification of the object category held in memory, regardless of the specific training scheme (PC: $t_{(1,23)}$ = 8.81, p < 0.001, η_p^2 = 1.80 or CP: $t_{(1,23)}$ = 9.04, p < 0.001, η_p^2 = 1.85). Moreover, we did not observe a difference in decoding performance between the two schemes (PC vs. CP: $t_{(1,23)}$ = 1.43, p = 0.167, η_p^2 = 0.29). These results indicate that during the delay prior to search, the representational pattern of the object category was similar regardless of the current or prospective status of the object.

Search 1: The prospective target can be decoded during the first search, but is different from its current counterpart

Next we were interested whether it was possible to successfully decode the prospective target while participants were searching for a different object. During the first search we observed clear decoding of the currently relevant object (compared against chance. 33%: $t_{(1,23)}$ = 11.57, p < 0.001, η_p^2 = 2.36, blue lines in Figure 2A), and this was stronger than for the prospectively relevant object (between-condition comparison: $t_{(1,23)}$ = 9.42, p < 0.001, η_p^2 = 1.92). This is to be expected as during the first search the current target category is actually presented on the screen, while the prospective target category is maintained in memory. Importantly, we were still able to also decode the prospective category during the first search (vs. 33%; $t_{(1,23)}$ = 1.90, p = 0.035, η_p^2 = 0.39 pink lines in Figure 2A). Thus, the prospectively relevant target category (dresser, cow, skate) can be successfully decoded when observers are performing a search for an unrelated target (the flower).

This then raises the question as to whether the re-emerging prospective representation resembles its counterpart when currently relevant. To assess this we used the cross-relevance decoding scheme. Remarkably, here we observed *below*-chance decoding performance during Search 1 (CP: $t_{(1,23)}$ = -4.79, p < 0.001, η_p^2 = -0.98, blue lines and PC: $t_{(1,23)}$ = -3.67, p = 0.001, η_p^2 = -0.75, pink lines). First, the reliable deviation from chance further confirms that information on the prospective memory was present in object-selective cortex during the first search. Second, the fact that decoding was below chance suggests that current and prospective representations of the same object category were represented through opposite multivariate patterns. As is shown in Supplementary Figure S1, this below-chance decoding did not coincide with the BOLD undershoot.

Finally, we wanted to assess how the dissociation between current and prospective representations generalized across the different phases of the trial. As Supplementary Figure S2 shows, the pattern of activity prior to search is very similar to that during search for currently relevant representations, whereas prospectively relevant representations during the first search are markedly dissimilar from the same categories during the delay period prior to search. Thus, while currently relevant representations remained constant from delay to search, the prospective representation was transformed from being similarly represented prior to search to being differently represented during search for the currently relevant item.

Search 2: Decoding of the first, now dropped target during the second search

Although not the primary goal of our study, we conducted the same analyses also for Search 2. As expected, here we saw the pattern reverse (see Figure 2A). In the within-relevance decoding scheme, we observed strong decoding of the category of the prospectively relevant target, which by now had become task-relevant (against chance, 33%: $t_{(1,23)}$ = 9.86, p < 0.001, η_p^2 = 2.01, pink lines). This decoding was stronger than for the previously current search target, which was now no longer relevant ($t_{(1,23)}$ = -7.51, p < 0.001, η_p^2 = -1.53). Nevertheless, and unexpectedly, we also observed above-chance decoding for this dropped first target during Search 2 ($t_{(1,23)}$ = 3.30, p = 0.002, η_p^2 = 0.67, blue lines). Note that this reflects classification of a target that is no *longer* relevant, whereas during Search 1 it reflected the target that was not *yet* relevant. Moreover, the cross-relevance decoding scheme also shows a pattern similar to what was observed during Search 1 (Figure 2B; and see also Figure S2 for the generalization across time). We found below-chance decoding in the same time range for both classification schemes (CP: $t_{(1,23)}$ = -4.65, p < 0.001, η_p^2 = -0.95, blue lines and PC: $t_{(1,23)}$ = -4.49, p < 0.001, η_p^2 = -0.92, pink lines). We will return to the re-emergence of the dropped target and the similarity in patterns to prospective memories in the General Discussion.

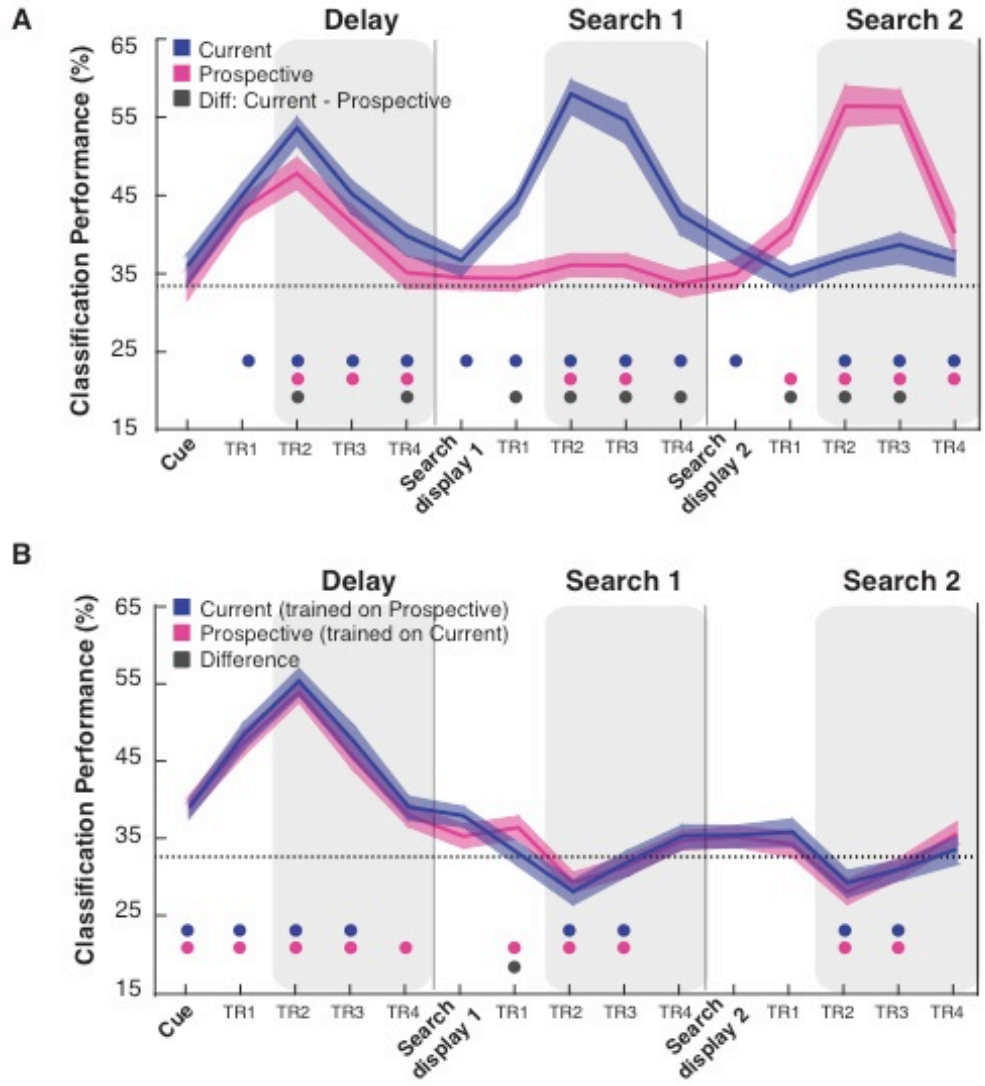


Figure 2. Within- relevance and cross-relevance object category decoding in pFs. (A) *Within-relevance decoding where the classifier was trained and tested within either current or prospective conditions. Object category decoding was higher for currently relevant objects (blue lines) than for prospectively relevant objects (pink lines) during the Delay prior to search and Search 1 and vice versa during Search 2. At the same time, the prospective target could reliably be decoded during the first search, while the dropped target could be decoded during the second search.* ***(B)*** *Cross-relevance decoding where the classifier was trained on current relevance, tested on prospective relevance, or vice versa. This resulted in above-chance decoding during the Delay prior to search (suggesting similar representations for current and prospective objects) but below-chance decoding during Search 1 and Search 2 (suggesting systematically different representations). Shaded blue and pink areas indicate between-subjects s.e.m. Circles on the bottom part of the figure indicate time points with significant above-chance information ($p < 0.05$). For the main analyses, data were averaged across the shaded time windows of three TRs each, N= 24 (see also Figure S3A).*

Representational dissimilarity analyses comparing current and prospective representations

To further elucidate the relationship between current and prospective representations, Figure 3A shows, for each interval of interest (Delay, Search 1, Search 2), the representational dissimilarity matrices (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008). Specifically, we cross-correlated the voxel response patterns for every possible pair of the 24 stimulus/relevance combinations (4 exemplars x 3 categories (Cow, Dresser and Skate) × 2 relevance (Current and Prospective; see Figure 3A right panel and Methods for further details). Figure 3B then shows multidimensional scaling (MDS) graphs of the same correlations to visualize the relationship between responses for each object category and relevance. The shorter the distance between categories the greater the representational similarity.

As can be seen from Figures 3A and 3B, throughout the course of the trial the neural representations moved from predominantly category space during the Delay period prior to search to predominantly relevance space during the two searches. Prior to search, objects grouped largely according to category, irrespective of relevance. This confirms that currently relevant and prospectively relevant objects were initially represented in similar ways in pFs. During Search 1, a clear relevance-driven distinction emerged between the neural object category representations. Note that this overall effect of relevance is probably driven by the fact that during search the currently relevant object type was presented in the display, while in the prospective condition the unrelated (flower) displays were presented. More interesting though is the finding that currently and prospectively relevant objects from the same category were represented as the most *dis*similar, as is illustrated by the these representations taking opposite corners in the MDS plot in Figure 3B. For example, while all four exemplars of the cow category clustered together when all current, or all prospective, current cows were most separated from prospective cows – to the extent that current cows were more closely represented to prospective skates and dressers than they were to prospective cows. The same pattern held for the two other categories.

To statistically test these effects, we computed the average dissimilarity between current and prospective objects, separately for when drawn from the same category (e.g. current cow versus prospective cow) and when drawn from a different category (e.g. current cow versus prospective skate/dresser) and used paired t-tests (N=24). Figure 3C shows these average same and different category dissimilarity values across relevance. During the Delay prior to the first search, as expected, *same* category representations were more similar than

different category representations across relevance ($t_{(1,23)}$ = -5.82, p < 0.001, η_p^2 = -1.28, as performed on Fisher-transformed 1-r values). In contrast, during Search 1, prospective targets differed most from current targets when they belonged to the same category, more so than when they belonged to different categories ($t_{(1,23)}$ = 3.06, p = 0.005, η_p^2 = 0.64). Likewise, during Search 2, dropped targets were less similar from relevant targets when they belonged to the same category, than when they belonged to different categories ($t_{(1,23)}$ = 4.75, p < 0.001, η_p^2 = 0.97). Thus these analyses statistically confirm what we can observe from the MDS plots, namely that current and prospective objects move from similar to opposite representations.

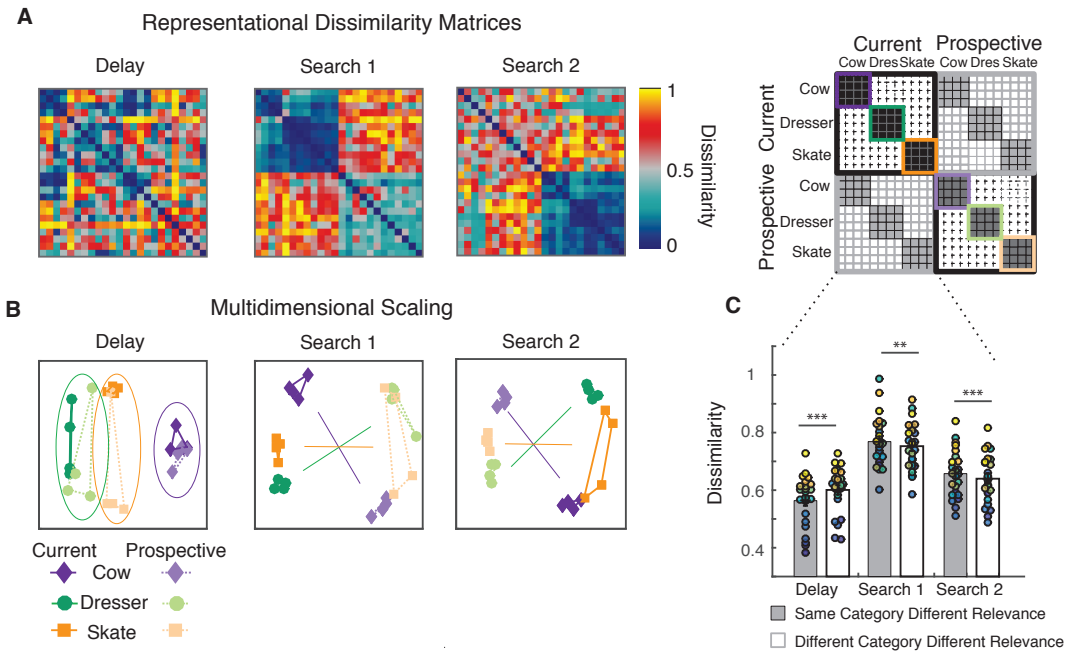


Figure 3. Representational dissimilarity of object representations in pFs. (A) *Representational dissimilarity matrices for the different target object categories during Delay, Search 1 and Search 2, as a function of relevance (current and prospective). Blue indicates that representations are more similar while red indicates more dissimilar* ***(B)*** *Multidimensional scaling plots of the same similarity values, for the same Delay, Search 1 and Search 2 intervals. The closer in space the more similar the neural representations. As a trial unfolds, object representations move from predominantly object category space (e.g. a cow) into predominantly relevance space (e.g. current target), where current and prospective targets of the same category are represented by partly opposite representational patterns.* ***(C)*** *Comparing dissimilarity between Current and Prospective items when they are drawn from the same category versus when they are drawn from different categories. Prior to search representations within the same category are more similar than different categories, but this reverses during the searches. Colored dots indicate individual participant data, **p < .01, ***p< .001, N = 24.*

How does voxel activity differ between current and prospective representations?

Next we wanted to investigate what turns a currently relevant into a prospectively relevant representation. To this end, we analyzed how voxel activity differs between these two relevance states as a function of the relative importance of the voxel in the representation, as indicated by the classifier weight. We first sorted the voxels according to their classifier weights obtained when the classifier was trained on *currently* relevant categories during the first search see Methods for details). Next, we correlated for each participant the classifier weights with the activation (t-value) of each condition (Current and Prospective (same and different category) using Spearman's Rho. We did this separately for each object category. For visualization purposes, we organized these sorted voxels into 30 bins in order to allow averaging over participants, as the number of voxels of the individually mapped pFs varied between participants. Figure 4 plots the voxels' signal strength – expressed as the *t*-value of the voxel's GLM fit – as a function of the voxels' binned classifier weights for each object category, plus the average across categories. As would be expected, voxel signal strength for currently relevant objects correlated positively with classifier weight, as the most important voxels tend to be those with the largest t-values (blue circles).

Figure 4 also shows the activity for *prospective* objects of a *different* category, still as a function of the classifier weight sorted on the basis of the currently relevant category (purple triangles). As expected, no relationship was observed here, since these activations belonged to different categories as well as a different state. Interestingly though, the voxel activity for prospective objects of the *same* category (again still sorted according to the classifier weights when current) anti-correlated with its current counterpart. Thus, voxel activity for prospective objects of the same category differ more from voxel activity of current objects than voxel activity of prospective objects of a different category. This pattern occurred for all three categories and for 23 out of the 24 participants, resulting in a highly reliable difference between the correlations (Spearman Rho's) for same and different prospective categories (paired t-test: $t_{(1,23)}$ = -7.60, p < 0.001, η_p^2 = 1.55). This analysis confirms that prospectively relevant voxel patterns anti-correlated with currently relevant voxel patterns of the same category, more so than of a different category. Taken together, the results indicate that the most active part of a representation becomes the least responsive, and vice versa, when moving from current to prospective.

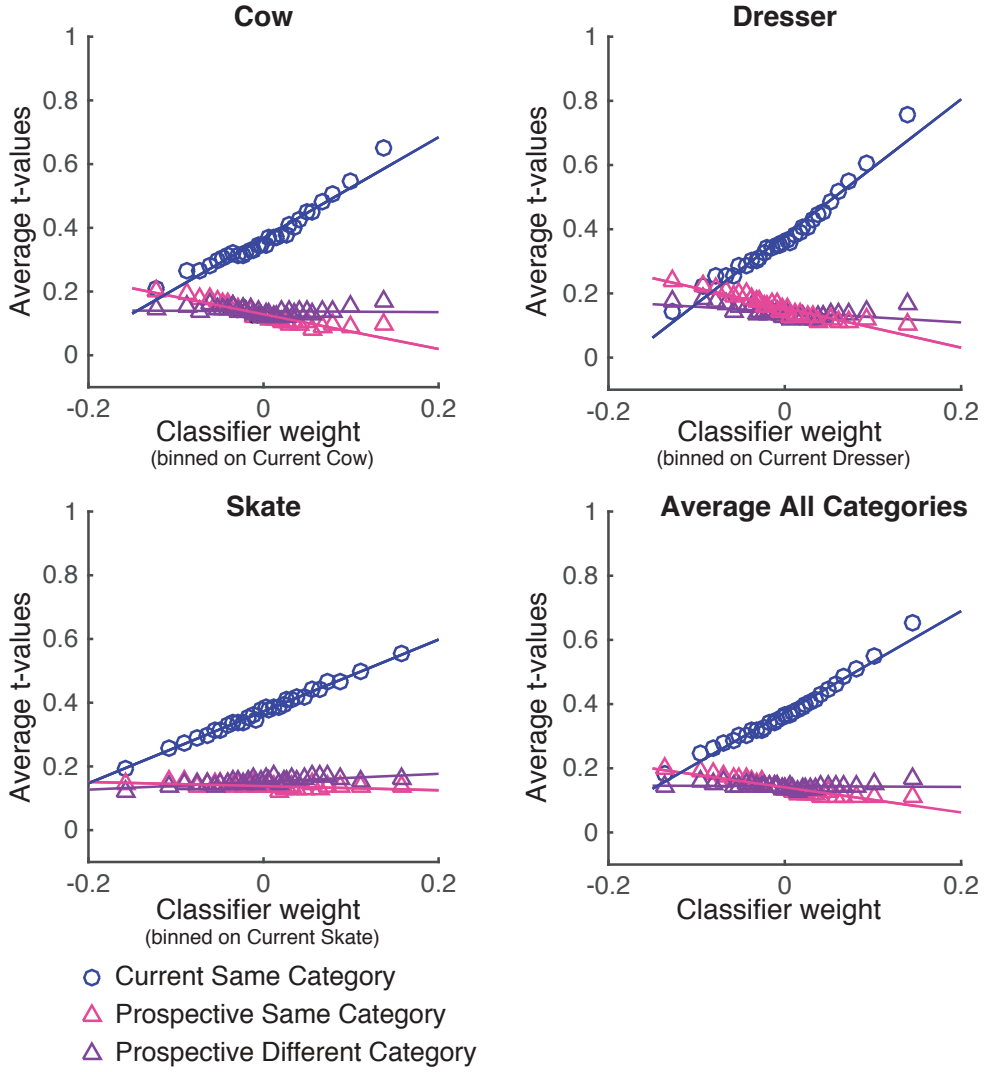


Figure 4. Voxel signal strength as a function of current voxel classifier weight. *The horizontal axis shows the relative importance of the voxels, expressed in classifier weights for currently relevant representations, while the vertical axis shows their signal strength, expressed in t-values, during Search 1. Voxel strength correlated positively with classifier weight for currently relevant objects (as searched for on the screen, blue circles), but the same voxels were overall less active and anti-correlated when objects were prospectively held in memory. Notably, activity associated with prospectively held objects of the same category differed more from their current counterpart (pink triangles) than prospectively maintained objects of a different category (purple triangles).*

Other regions of interest

Finally, to see if our results were specific to object-selective cortex, we ran the same analyses on other regions of interest (ROIs) that have been implicated in working memory and visual search: rostral middle frontal cortex (RMF), the intra parietal sulcus (IPS) and early visual cortex (VC). In addition we performed a classification analyses based on relevance alone, irrespective of object category. The results are shown in Supplementary Figure S3. We found that object category information was stronger for pFs, visual cortex and IPS than for

RMF, whereas the impact of relevance (Current vs Prospective) was apparent in all ROIs throughout the trial. The findings provide support for the idea that parietal and prefrontal regions can be sensitive to specific memory content, but may have as their primary role planning and overall goal maintenance, while stimulus-specific representations are maintained more strongly posteriorly in sensory areas, including visual cortex (Bugatus, Weiner, & Grill-Spector, 2017; Hebart, Bankson, Harel, Baker, & Cichy, 2018; Lee et al., 2013; Postle, 2006).

Discussion

It has been shown that the content of working memory can be decoded from multivariate patterns of voxel activity when observers are remembering an item for a single task (Albers, Kok, Toni, Dijkerman, & de Lange, 2013; Harrison & Tong, 2009; Lewis-Peacock & Postle, 2012; Serences, Ester, Vogel, & Awh, 2009). Furthermore, working memory representations have been shown to adapt to the specific task goal, as the representation of the same object changes depending on the nature of the upcoming test (Lee et al., 2013; Myers, Stokes, & Nobre, 2017). Here we provide further evidence for the flexibility of working memory by showing how it distinguishes between current and future task goals, as representations adapt to the order in which they are required in multiple task sequences.

In line with earlier work in different task and stimulus domains (LaRocque et al., 2013; 2017; Lewis-Peacock & Postle, 2012; Wolff et al., 2017), we found that objects required for an upcoming search task are represented more strongly than when the same objects are only used prospectively, for a subsequent search task, as was indicated by stronger classification performance in object-selective cortex prior to the first search. Furthermore, corroborating findings by Rose et al., (2016) and Wolff et al., (2017), we found that the prospectively relevant memory could be reconstructed during task-irrelevant stimulation, here during the first search for an unrelated stimulus. This was shown in two ways: First, we found above-chance classification of the prospective item during the first search. Second, in a cross-relevance training scheme, where the classifier was trained when the item was current and tested when it was prospective (or vice versa), we also found that decoding reliably differed from chance – but now in a negative direction.

Most importantly, our results are the first to reveal a direct relationship between currently relevant object representations on the one hand, and prospective representations on the other. Prior to the first search these types of representation were very similar, as cross-relevance decoding (training on one status while testing on the other) showed above-

chance classification performance. However, this relationship reversed during search, where the reconstructed prospective memory representations of objects now proved *dis*similar from their current counterparts. Importantly, they differed in a systematic manner, to the extent that prospective representations were even more dissimilar from current representations of the same object category than representations of a different object category, and were characterized by an inverse correlation with current representations. Thus, prospective targets may be dissociated from current targets in two ways. First, they appear distinct in that current representations are activity-based, whereas prospective representations are responsivity-based. Second, prospective targets are represented through a pattern of responsivity opposite to that of current target activity, where the most active part becomes the least responsive and vice versa.

Imaging studies of visual attention have demonstrated that attending to task-relevant objects at the expense of irrelevant objects leads to the transformation of representational space, by relatively enhancing target-related distributed activity (Reddy, Kanwisher, & VanRullen, 2009) and by recruiting additional resources as neurons shift their tuning towards the attended object category (Çukur, Nishimoto, Huth, & Gallant, 2013; Nastase et al., 2017). Here, such relative enhancement can explain the pattern of results during the delay period prior to the first search task, where we found a difference in representational strength between current and prospective memories, while their representations remained similar. However, the pattern during search suggests that in addition to any enhancement of the currently relevant representation, prospective representations are being separated by inversely transforming the responsivity of the neuronal populations.

A crucial question that our data does not answer is what the exact mechanism is behind this transformation. First note that we found evidence for an inversion of the representation both when the object was *temporarily* irrelevant – that is, it had a prospective status during the first search – and when it was *no longer* relevant – that is, during the second search, when the first target could be dropped. This suggests a shared mechanism for making a memory irrelevant, whether temporarily or for the remainder of the trial. One possibility is the involvement of an active mechanism of cognitive control, which attempts to maximally dissociate prospectively relevant from currently relevant representations in order to prevent task interference. Such control mechanisms might be exerted through feedback connections emanating from frontal areas central to counteracting unwanted or task-irrelevant information (Anderson et al., 2004; Banich,

Mackiewicz Seghete, Depue, & Burgess, 2015; de Vries, van Driel, Karacaoglu, & Olivers, n.d.; Depue, Curran, & Banich, 2007; Reeder, Olivers, & Pollmann, 2017). Interestingly, an earlier study of memory retrieval has shown suppression of voxel patterns in ventral object-related cortex which were associated with task-irrelevant memories of learned object pictures, leading to comparable patterns of representational dissimilarity as here (Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). Initial evidence for the suppression of temporarily irrelevant items also comes from a study by Peters, Roelfsema, & Goebel (2012), who used a similar task design as ours. They asked observers to consecutively look for a particular house target and a particular face target (or vice versa) in rapid streams of house/face distractors. They found the overall BOLD signal to be reduced in either house or face selective areas in response to house/face stimuli when the respective target was prospectively relevant. Here we show how changing task relevance within working memory specifically affects the cortical pattern of activation within memory while observers perform a different search task.

However, a distinct alternative possibility is that local, and arguably more passive mechanisms cause the change in responsiveness. One candidate is neural adaptation, as neural firing activity wears out with prolonged presentation (Henson & Rugg, 2003; Larsson & Smith, 2012; Vautin & Berkley, 1977). This may include changes in the membrane potentials of the previously firing neurons, as was proposed by Stokes (2015) as one possibility for latent storage. In our experiment, attending to the cued target object prior to search would then first be accompanied by the associated pattern of activity, which then automatically results in adaptation being the strongest for the most active neurons after activity has been switched off. As a consequence, the same neurons would be less responsive when new stimulation arrives, resulting in the pattern observed here. Such adaptation in responsivity might also explain the earlier demonstrations of memory reconstruction by Rose et al. (2016) and Wolff et al. (2017). Importantly, for such adaptation to have a functional role in prospective working memory maintenance, we must assume that it can serve later retrieval – that is, the brain has ways of reading out the information stored in the changed responsivity when the representation becomes relevant again (Meyer & Rust, 2018; Turk-Browne, Yi, & Chun, 2006; Ward, Chun, & Kuhl, 2013).

Although the underlying mechanism remains unknown, we believe the results have important implications for theories of prospective memory storage in working memory. First, the fact that prospectively relevant objects could be decoded from the same regions of interest as the currently relevant objects indicates that the different memory states do not necessarily rely on different brain areas. The successful cross-relevance decoding, where we

trained the classifier on one state and tested on another, further confirms this. Prospective memories might have been stored at a more abstract or verbal level. However, our analysis of other brain regions, including frontal regions associated with more linguistic representations did not reveal category specificity that was specific to just prospective memories. Of course, our analyses of these regions were coarse and do not exclude the possibility that other brain areas are involved in representing either current or prospective information (cf. Christophel, Iamshchinina, Yan, Allefeld, & Haynes, 2018).

Second, the idea that prospectively relevant memories are stored in an activity-silent format has recently been debated by Schneegans and Bays (2017) on the basis of the argument that existing data can also be explained by a simpler model which assumes that temporarily irrelevant memories are represented through the same activity as relevant memories, but in a weaker form. Schneegans and Bays (2017) argued specifically against a study by Sprague, Ester, & Serences, 2016, which indeed showed clear remnants of activity for representations that were assumed to be partly latent. But even when the data reveals no such activity this may only reflect the limited sensitivity of the measure at hand. The reduced activity account is partially supported by our data. We found that during the delay period prior to the first search task, before the evidence for the prospective item diminished to baseline levels, current and prospective representations were highly similar, as evidenced by a strong correlation and successful cross-relevance classification of the current and prospective representations. However, the reduced activity account does not explain that current and prospective representational patterns were very dissimilar during the search. In fact, the partial anti-correlation indicates suppression rather than activation of the relevant voxels.

Instead, the emergence of the prospective memory that we found here during the first search fits best with a change in responsivity, resulting in an activity-silent representation. The fact that it was necessary to add activity to the system for the prospective memory to emerge – here in the form of unrelated visual search displays, is already testament to this. Importantly, the current data puts limits on the potential mechanisms by which the responsivity changes. A frequent hypothesis is that prospectively relevant representations are stored through temporary potentiation of the relevant connections, through synaptic weight changes. Such short-term potentiation predicts that what was strongly activated during encoding, becomes more strongly connected, and thus more responsive, when prospective. This is not what we observed in our experiment. We found the opposite: What was strongly activated when current, became more strongly

suppressed when prospective and vice versa. This goes against a simple short-term potentiation account of activity-silent representations in working memory.

In conclusion, we find evidence that, in trying to separate current and prospective goals in visual search, the brain stores representations within the same neuronal ensembles, but through opposite representational patterns.

Methods

Participants

Twenty-four subjects (8 males, *M* = 26.74 years of age, SD = 3.21 years) participated in the study. We obtained written informed consent from each participant before experimentation. Participants had normal or corrected-to-normal vision. The experiment was approved by the Ethical Committee of the Faculty of Social and Behavioural Sciences, University of Amsterdam and conformed to the Declaration of Helsinki.

Task and Stimuli

On each trial, participants performed two consecutive visual search tasks of real-world objects. The object of interest (cow, dresser or skate) consisted of real-world greyscale photographs, selected out of 4 exemplars, These categories were selected to have maximal dissimilarity in representational space (see Harel et al., 2014). The object of interest (cow, dresser or skate) was to be searched for first, or second – thus making it currently or prospectively relevant. To maximize the chances of decoding the target of interest (whether current or prospective), and to limit the working memory load, the remaining search task always involved the same 'daisy' flower target (only 1 exemplar).

As can be seen in Figure 1, each trial started with a fixation followed by the sequential presentation of two memory items (object of interest (cow, dresser or skate) and the daisy, 2.4° visual angle) each presented for 750 ms with a 500 ms fixation in between. After a fixation of 500 ms a cue, either a 1 or a 2 was presented indicating the search order in which the memory items needed to be searched for in two subsequent search tasks. Thus, participants either had to search for the objects of interest first (current) and then the daisy in the second search (referred to as Current condition) or the daisy had to be searched for first and the object of interest second (referred to as Prospective condition). Both relevance conditions (Current and Prospective), order of the memory items as well as the cue was counterbalanced across trials. The cue was followed by an 8 second delay with a fixation dot in the middle of the screen ('Delay') after which the first search display was presented. The search display consisted of 6 different exemplars (2.4° visual angle) of the same category as the cued memory item and could either contain the remembered 'Current' object ('Present') or not ('Absent'). Participants had to indicate through button presses with their left and right hand whether the memory item was present of absent. The distractors in the search displays were randomly placed among a radius of (7.4° visual angle). The search display was presented for two seconds and participants had to respond within these two seconds. After the first search display another eight seconds blank delay period followed ('Search 1')

followed by the second search display now depicting exemplars from the uncued object category. This was again followed by an eight second inter trial interval (ITI) ('Search 2'). After completion of the first search task, observers had to turn to the prospective item, and indicate its presence or absence in the second search display. Participants received feedback (for 400 ms) on their performance for search tasks either 'correct', 'incorrect' or 'missed' (if not responded within the 2 seconds of the search displays) after each trial within the ITI. At the end of each run the percentage correct and average reaction times were presented to the participants separate for the flower search task and other objects search.

The stimuli were back-projected on a 61 x 36 cm LCD screen (1920 x 1080 pixels) using Presentation (Neurobehavioral Systems, Albany, CA, USA) and viewed through a mirror attached to the head coil. Eye tracking data (EyeLink 1000, SR Research, Canada) were recorded to ensure participants were awake and attending the stimuli. The main experiment consisted of 8 runs with 12 trials each (96 trials in total). Each experimental run had a duration of ~ 7 minutes. The total duration of a session was ~1.5 hours (including the structural scan (6 minutes) and mapper run (7 minutes), see below).

Regions of Interest: object-selective cortex mapper (pFs)

At the end of each session we independently mapped the region of interest as the region that responded more strongly to intact vs. scrambled objects (Malach et al., 1995), within an anatomical mask of the temporal occipital fusiform cortex (from the Harvard-Oxford Cortical Structural Atlas of the FSL package; see Figure S3A). We used the same images and object categories as in our experimental task (cow, skate, dresser and flower). This localized object-selective region of interest corresponded to the posterior fusiform part of lateral occipital cortex (pFs). Stimuli were presented for 250 ms and consisted of 48 intact objects (12 of each object category) and 48 scrambled objects (12 of each object category) that were presented in separate blocks for each object category (24 in total) with fixation block intermixed (seven in total). The mapper run had a duration of ~ 7 minutes. Participants were asked to push a button when two consecutive images were identical (same exemplar) to ensure attention. The same fMRI preprocessing steps as described for the experimental task were performed for this mapper. For two participants the data recorded from this mapper was not usable, therefore we used the anatomical mask only for these participants.

fMRI Acquisition

Scanning was done on a 3T Philips Achieva TX MRI scanner with a 32-elements head coil. In the middle of each session (after 4 runs) a high-resolution 3DT1-weighted anatomical image (TR, 8.35 ms; TE, 3.83 ms; FOV, 240 × 220 × 188, 1 mm^3 voxel size) was recorded for every participant (duration 6 minutes).

During the experimental task and objective-selective cortex functional localizer, blood oxygenation level dependent (BOLD)-MRI was recorded using Echo Planar Imaging (EPI) (TR 2000 ms, TE 27.62 ms, FA 76.1, 36 slices with ascending acquisition, voxel size 3 mm^3, slice gap 0.3 mm, FOV 240 x 118.5 x 240).

fMRI Data analysis

fMRI Preprocessing

Anatomical T1 scans were automatically segmented using the Freesurfer package (http://surfer.nmr.mgh.harvard.edu/; Dale, Fischl, & Sereno, 1999). BOLD-MRI data was registered to the subject specific T1 scan using boundary based registration (Greve & Fischl, 2009). The subject-specific T1 scan was registered to the MNI brain using FMRIB's Nonlinear Image Registration Tool (FNIRT). For the functional imaging data we used FEAT version 5, part of FSL (Oxford Centre for Functional MRI of the Brain (FMRIB) Software Library; www.fmrib.ox.ac.uk/fsl; Smith et al., 2004). Preprocessing steps consisted of motion correction, brain extraction, slice-time correction, alignment, and high-pass filtering (cutoff 100 s). For each subject and each trial a general linear model (GLM) was fitted to the data, whereby every TR (2 seconds each) was taken as a regression variable. We derived the t-value of each voxel for each of the fifteen TRs 2-30 s after the start of each trial. We used FMRIB's Improved Linear Model (FILM) (Woolrich, Ripley, Brady, & Smith, 2001) for the time-series statistical analysis. The data was further analyzed in Matlab (The MathWorks, Natick, MA, USA). For every participant, every run, every experimental condition (status (Current and Prospective) x category exemplar (Cow, Dresser and Skate, 4 exemplars of each) and for each TR, we created a vector containing the *t*-value per voxel in our regions of interest (see below). T-values for each predictor were computed by dividing the beta-weight by the standard error. That vector comprised the spatial pattern of activity evoked at that time point (TR) for that experimental condition in a specific region of interest.

Within-relevance and Cross-relevance Object Category Decoding

Next, we used these multi-voxel patterns to answer the question whether Relevance (current or prospective) affected the neural category representations. To determine this we used the Princeton Multi-Voxel Pattern Analysis toolbox (available at https://github.com/princetonuniversity/princeton-mvpa-toolbox, see Detre et al. 2006). To examine whether current and prospective items evoked a distinct pattern of activity in the regions of interest, a single class logistic regression classifier was trained to distinguish each object category (cow, dresser and skate). Logistic regression computes a weighted combination of voxel activity values, and it adjusts the (per-voxel) regression weights to minimize the discrepancy between the predicted output value and the correct output value. The maximum number of iterations used by the iteratively-reweighted least squares (IRLS) algorithm was set to 5000. Classifier performance was evaluated with a leave one run out cross validation procedure. This involved training a single class logistic regression classifier to learn a mapping between the neural patterns and the corresponding category labels for all but one run, and then using the trained classifier to predict the category of stimuli from the test patterns in the remaining run. For each iteration we trained the classifier on seven runs and tested on the remaining run for each ROI

We investigated category decoding (Cow, Dresser and Skate) both within Current and Prospective relevance (*within-relevance decoding*) and between relevance conditions (*cross-relevance decoding*) for each time point (TR) in the trial separately. We trained and tested the classifier on the same relevance (Current or Prospective) as well as across relevance conditions - i.e. trained when the category was a Current item and tested when the category was a Prospective item ('PC') and vice versa ('CP'). This yielded a classification score (percentage correct) per subject for every condition (Category) and time point (TR) depending on the status of the object. Note that here chance decoding was 33.33% since we had three object categories (Cow, Dresser and Skate). All statistical comparisons are based on two-tailed tests, except for the comparison against chance in the within-relevance coding scheme as there decoding cannot go below chance (cf. Christophel etal., 2018). All statistical analyses were performed using SPSS 17.0 (IBM, Armonk, USA)

Representational dissimilarity Analysis

For each TR we created a representational dissimilarity matrix (RDM) (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013). Each cell of the matrix represents a 1-rho (Spearman correlation) of the activity patterns of two individual exemplars. In this experiment our

RDMs consisted of 24x24, 4 unique exemplars per category (Cow, Dresser and Skate) and 2 different relevance levels (Current and Prospective). The RDMs of each run were averaged to obtain one RDM per TR. We further averaged across the three TRs for each interval of interest in the trial (Delay, Search 1 and Search 2). For visualization purposes we transformed the RDM by replacing each element by its rank in the distribution of all its elements (scaled between 0 to 1). In addition, we used multidimensional scaling (MDS) plots wherein the distance between points reflects the dissimilarity in their neural patterns of response. To compute the interaction between Relevance and Category over the course of the trial we calculated the dissimilarity for the between Relevance (Current vs Prospective) and Category (same (Cow/Cow, Dresser/Dresser and Skate/Skate) vs different (Cow/Dresser, Dresser/Skate, Skate/Dresser)) by averaging the cells within each class. We calculated this for every TR separately, and then averaged those across the three TRs in the predetermined intervals (Delay, Search 1 and Search 2).

Correlating classifier weights with voxel signal strength

For each participant, we extracted for each voxel in the pFs the classifier weight for each category (cow, dresser and skate), obtained from training the classifier on Current object categories during the first search (Search1_Tr2). Second, we correlated these weights with the corresponding t-values (signal strength) at Search1_Tr2 separately for each condition (current and prospective) and both within and between categories for each participant as well as for each object category. For example, we extracted the voxel classifier weights for cows and correlated these with the voxel signal strength for currently relevant cows, prospectively relevant cows (same category) and prospectively relevant dressers and skates (different category). This was then averaged across the different objects. Note that for the classifier to be biased against prospective representations of a particular category, it only needs to prefer one of the other categories (other than the trained one), which happens to be more similar to the trained representation. So we compared the same category correlation to this preferred different category correlation using a paired t-test (two-sided) on the Fisher transformed correlations. However, all results for this analysis also hold when we take the average rather than the preferred of the two alternative different categories.

For visualization purposes (see Figure 4), because the number of voxels varied across participants, we binned the classifier weights for each participant using 30 bins and then averaged over participants for each bin.

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared Representations for Working Memory and Mental Imagery in Early Visual Cortex. *Current Biology*, *23*(15), 1427–1431. http://doi.org/10.1016/j.cub.2013.05.065

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., et al. (2004). Neural systems underlying the suppression of unwanted memories. *Science*, *303*(5655), 232–235. http://doi.org/10.1126/science.1089504

Arbuckle, S. A., Yokoi, A., Pruszynski, J. A., & Diedrichsen, J. (2018). Stability of representational geometry across a wide range of fMRI activity levels. bioRxiv, http://doi.org/10.1101/266585

Banich, M. T., Mackiewicz Seghete, K. L., Depue, B. E., & Burgess, G. C. (2015). Multiple modes of clearing one's mind of current thoughts: overlapping and distinct neural systems. *Neuropsychologia*, *69*, 105–117. http://doi.org/10.1016/j.neuropsychologia.2015.01.039

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, *25*, 20–24. http://doi.org/10.1016/j.conb.2013.10.008

Bugatus, L., Weiner, K. S., & Grill-Spector, K. (2017). Task alters category representations in prefrontal but not high-level visual cortex. *Neuroimage*, *155*, 437–449. http://doi.org/10.1016/j.neuroimage.2017.03.062

Carlisle, N. B., & Woodman, G. F. (2011). Automatic and strategic effects in the guidance of attention by working memory representations. *Acta Psychologica*, *137*(2), 217–225. http://doi.org/10.1016/j.actpsy.2010.06.012

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. http://doi.org/10.1038/nn.3381

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*(1), 115–142. http://doi.org/10.1146/annurev-psych-010814-015031

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis - I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194. http://doi.org/10.1006/nimg.1998.0395

de Vries, I. E. J., van Driel, J., Karacaoglu, M., & Olivers, C. N. L. (n.d.). Priority switches in visual working memory are supported by frontal delta and posterior alpha interactions. *(Under Review)*.

Depue, B. E., Curran, T., & Banich, M. T. (2007). Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science*, *317*(5835), 215–219. http://doi.org/10.1126/science.1139560

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980. http://doi.org/10.1016/j.neuroimage.2006.01.021

Detre, G. J., Polyn, S. M., Takerkart, S., Natu, V. S., Benharrosh, M. S., Singer, B. D., ... & Norman, K. A. (2006). The multi-voxel pattern analysis (MVPA). In *12th Meeting of the Organization of Human Brain Mapping, Florence, Italy*.

Downing, P. E., & Dodds, C. M. (2003). Competition in visual working memory for control of search. *Perception*, *32*, 92–93.

Erickson, M. A., Maramara, L. A., & Lisman, J. (2010). A Single Brief Burst Induces GluR1-dependent Associative Short-term Potentiation: A Potential Mechanism for Short-term Memory. *Journal of Cognitive Neuroscience*, *22*(11), 2530–2540. http://doi.org/10.1162/jocn.2009.21375

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-

based registration. *Neuroimage*, *48*(1), 63–72. http://doi.org/10.1016/j.neuroimage.2009.06.060

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(10), E962–71. http://doi.org/10.1073/pnas.1312567111

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. http://doi.org/10.1038/nature07832

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife*, *7*. http://doi.org/10.7554/eLife.32816

Henson, R., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, *41*(3), 263–270.

Houtkamp, R., & Roelfsema, P. R. (2006). The effect of items in working memory on the deployment of attention and the eyes during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 423–442. http://doi.org/10.1037/0096-1523.32.2.423

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. http://doi.org/10.1016/j.tics.2014.01.002

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49. http://doi.org/10.1016/j.tics.2012.10.011

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. http://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. http://doi.org/10.1016/j.neuron.2008.10.043

LaRocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, *8*. http://doi.org/10.3389/fnhum.2014.00005

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142. http://doi.org/10.1162/jocn_a_00305

LaRocque, J. J., Riggall, A. C., Emrich, S. M., & Postle, B. R. (2017). Within-Category Decoding of Information in Different Attentional States in Short-Term Memory. *Cerebral Cortex*, *27*(10), 4881–4890. http://doi.org/10.1093/cercor/bhw283

Larsson, J., & Smith, A. T. (2012). fMRI Repetition Suppression: Neuronal Adaptation or Stimulus Expectation? *Cerebral Cortex*, *22*(3), 567–576. http://doi.org/10.1093/cercor/bhr119

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997–999. http://doi.org/10.1038/nn.3452

Lewis-Peacock, J. A., & Postle, B. R. (2012). Neuropsychologia. *Neuropsychologia*, *50*(4), 470–478. http://doi.org/10.1016/j.neuropsychologia.2011.11.006

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. http://doi.org/10.1162/jocn_a_00140

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-Related Activity Revealed by Functional Magnetic Resonance Imaging in Human Occipital Cortex. *Proceedings of the National Academy of Sciences of the United States*

of America, *92*(18), 8135–8139. http://doi.org/10.2307/2368011?ref=search-gateway:ae8ad9f1e027b5b5af79d6209b7141b9

Mallett, R., & Lewis-Peacock, J. A. (2018). Behavioral decoding of working memory items inside and outside the focus of attention. *Annals of the New York Academy of Sciences*, *12*, 342. http://doi.org/10.1162/jocn_a_01180

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. http://doi.org/10.1016/j.jneumeth.2007.03.024

Meyer, T., & Rust, N. (2018). Single-exposure visual memory judgments are reflected in inferotemporal cortex. *eLife*, *7*. http://doi.org/10.7554/eLife.32259

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543–1546. http://doi.org/10.1126/science.1150769

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*, *21*(6), 449–461. http://doi.org/10.1016/j.tics.2017.03.010

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Castello, M. V. D. O., et al. (2017). Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. *Cerebral Cortex*, *27*(8), 4277–4291. http://doi.org/10.1093/cercor/bhx138

Olivers, C. N. L., & Eimer, M. (2011). On the difference between working memory and attentional set. *Neuropsychologia*, *49*(6), 1553–1558. http://doi.org/10.1016/j.neuropsychologia.2010.11.033

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334.

Peters, J. C., Roelfsema, P. R., & Goebel, R. (2012). Task-Relevant and Accessory Items in Working Memory Have Opposite Effects on Activity in Extrastriate Cortex. *The Journal of Neuroscience*, *32*(47), 17003–17011. http://doi.org/10.1523/JNEUROSCI.0591-12.2012

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. http://doi.org/10.1016/j.neuroscience.2005.06.005

Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21447–21452. http://doi.org/10.1073/pnas.0907330106

Reeder, R. R., Olivers, C. N. L., & Pollmann, S. (2017). Cortical evidence for negative search templates. *Visual Cognition*, *25*(1-3), 278–290. http://doi.org/10.1016/0306-4522(83)90265-8

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. http://doi.org/10.1126/science.aah7011

Schneegans, S., & Bays, P. M. (2017). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience*, *29*(12), 1977–1994. http://doi.org/10.1038/nature04262

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, *20*(2), 207–214. http://doi.org/10.1111/j.1467-9280.2009.02276.x

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208–S219. http://doi.org/10.1016/j.neuroimage.2004.07.051

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory

Representations in Human Cortex. *Neuron*, *91*(3), 694–707. http://doi.org/10.1016/j.neuron.2016.07.006

Stokes, M. G. (2015). “Activity-silent” working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. http://doi.org/10.1016/j.tics.2015.05.004

Turk-Browne, N. B., Yi, D.-J., & Chun, M. M. (2006). Linking Implicit and Explicit Memory: Common Encoding Factors and Shared Representations. *Neuron*, *49*(6), 917–927. http://doi.org/10.1016/j.neuron.2006.01.030

van Loon, A. M., Olmos Solis, K., & Olivers, C. N. L. (2017). Subtle eye movement metrics reveal task-relevant representations prior to visual search. *Journal of Vision*, *17*(6), 13. http://doi.org/10.1167/17.6.13

Vautin, & Berkley, M. A. (1977). Responses of Single Cells in Cat Visual-Cortex to Prolonged Stimulus Movement - Neural Correlates of Visual Aftereffects. *Journal of Neurophysiology*, *40*(5), 1051–1065. http://doi.org/10.1152/jn.1977.40.5.1051

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. http://doi.org/10.1093/cercor/bhu277

Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *The Journal of Neuroscience*, *33*(37), 14749–14757. http://doi.org/10.1523/JNEUROSCI.4889-12.2013

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589. http://doi.org/10.1038/nn.3973

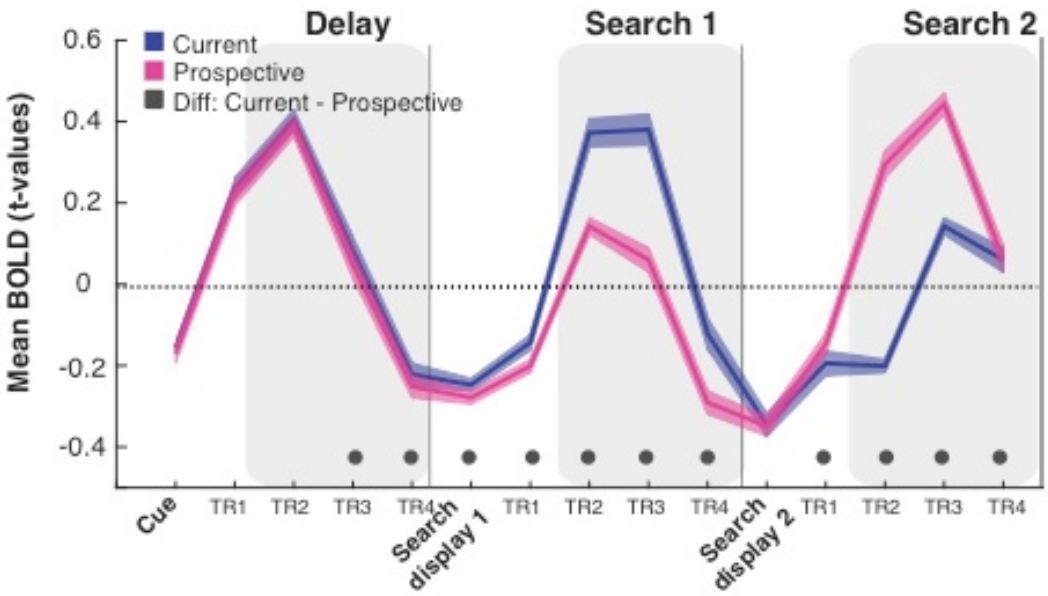
Wolff, M. J., Jochim, J., Akyurek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–. http://doi.org/10.1038/nn.4546

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, *14*(6), 1370–1386. http://doi.org/10.1006/nimg.2001.0931

Supplementary material and Figures

Mean BOLD response

Figure S1 shows the mean BOLD response in area pFs for the current and prospective conditions, averaged for the same individual ROIs as used in the MVPA. From this there is no evidence that the sharp but temporary negative classification is caused by the BOLD undershoot. Recent work furthermore suggests that multivariate representational patterns are stable across different levels of BOLD response (Arbuckle, Yokoi, Pruszynski, & Diedrichsen, 2018).

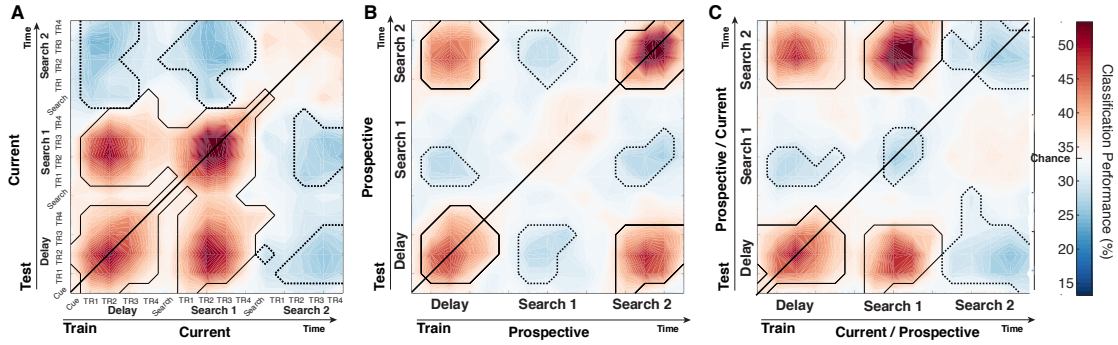


Supplementary **Figure S1: Mean BOLD response.** *During the Delay there was a small difference in the BOLD response magnitude depending on whether the category was currently or prospectively relevant ($t_{(1,23)}$ = 2.15, p = 0.0427, η_p^2 = 0.44). A stronger difference came apparent for Search 1 ($t_{(1,23)}$ = 14.46, p < 0.001, η_p^2 = 1.77) and Search 2 ($t_{(1,23)}$ = -13.08, p < 0.001, η_p^2 = -2.67), where the displays containing our object categories of interest elicited higher responses than those with flowers, probably due to higher task-demand since the flower search task was easier due to the repetition of the Daisy. Shaded areas indicate between-subjects s.e.m. Circles on the bottom part of the figure indicate time points with significant differences between Current and Prospective (p < 0.05), N=24.*

Cross-temporal generalization of object category decoding

The main analyses are based on decoding performance where training and testing occurred separately for each TR. To examine whether the neural representations of the different object categories for current and prospective states were also related across time, we tested for cross-temporal generalization of decoding accuracy (see King & Dehaene, 2014), by training the classifier on each of the TRs and then testing it on all other TRs in the trial. This was then repeated for all TRs, creating a two-dimensional matrix of cross-temporal object category decoding (with no additional smoothing). Time windows of significant decoding were identified using 2-dimensional cluster-based permutation testing (i.e., across both time axes) with cluster correction ($p = 0.05$ and 10.000 iterations) to statistically compare the object category decoding with chance (33.33%) (Maris & Oostenveld, 2007) using and Matlab (The MathWorks, Natick, MA, USA). As a result, we were able to assess the temporal stability of object category decoding and to test whether encoding and maintenance of the object (Delay) was similar to searching for an object (Search 1 and Search 2).

Figure S2 shows the resulting temporal generalization matrices, with red indicating above-chance decoding performance and blue indicating below-chance decoding performance. Note that the pattern on the diagonal reflects the classification per TR as in Figure 2. Of additional interest here are the significant off-diagonal clusters as they indicate a generalized representation across time. These clusters show that target representations maintained during maintenance prior to the first search were similar to the pattern of activity that is observed when subjects are looking at these object categories during search (i.e. the first search when current; see the red off-diagonal clusters in Figure S2A; and the second search when prospective; see the red off-diagonal clusters in Figure S2B). Figure S2C shows the temporal generalization for the cross-relevance decoding scheme (averaged across PC and CP schemes). Notable here is that the same target representations as maintained during maintenance prior to the first search are *dis*similar to the prospective representations during the first search, and the dropped representations during the second search, as indicated by the blue off-diagonal clusters.



Supplementary Figure S2. Cross-temporal generalization matrices for object category decoding as a function of Relevance, with (A) *the Current condition,* ***(B)*** *the Prospective condition, and* ***(C)*** *cross-relevance classification (averaged for training on current and testing on prospective with the transpose of the reverse training scheme). We observed that target representations for the objects of interest maintained during the Delay period prior to the first search generalized to the search displays that contained that target (i.e. the first search when current, or the second search when prospective), as shown by reliable off-diagonal red clusters. In contrast, off-diagonal blue clusters emerge when the object trained during the delay is prospective during Search 1, or dropped during Search 2, indicating anti-correlated representations. Outlines indicate cluster-based permutation tests with p < .05, positive clusters (solid lines), negative clusters (dotted lines), N=24.*

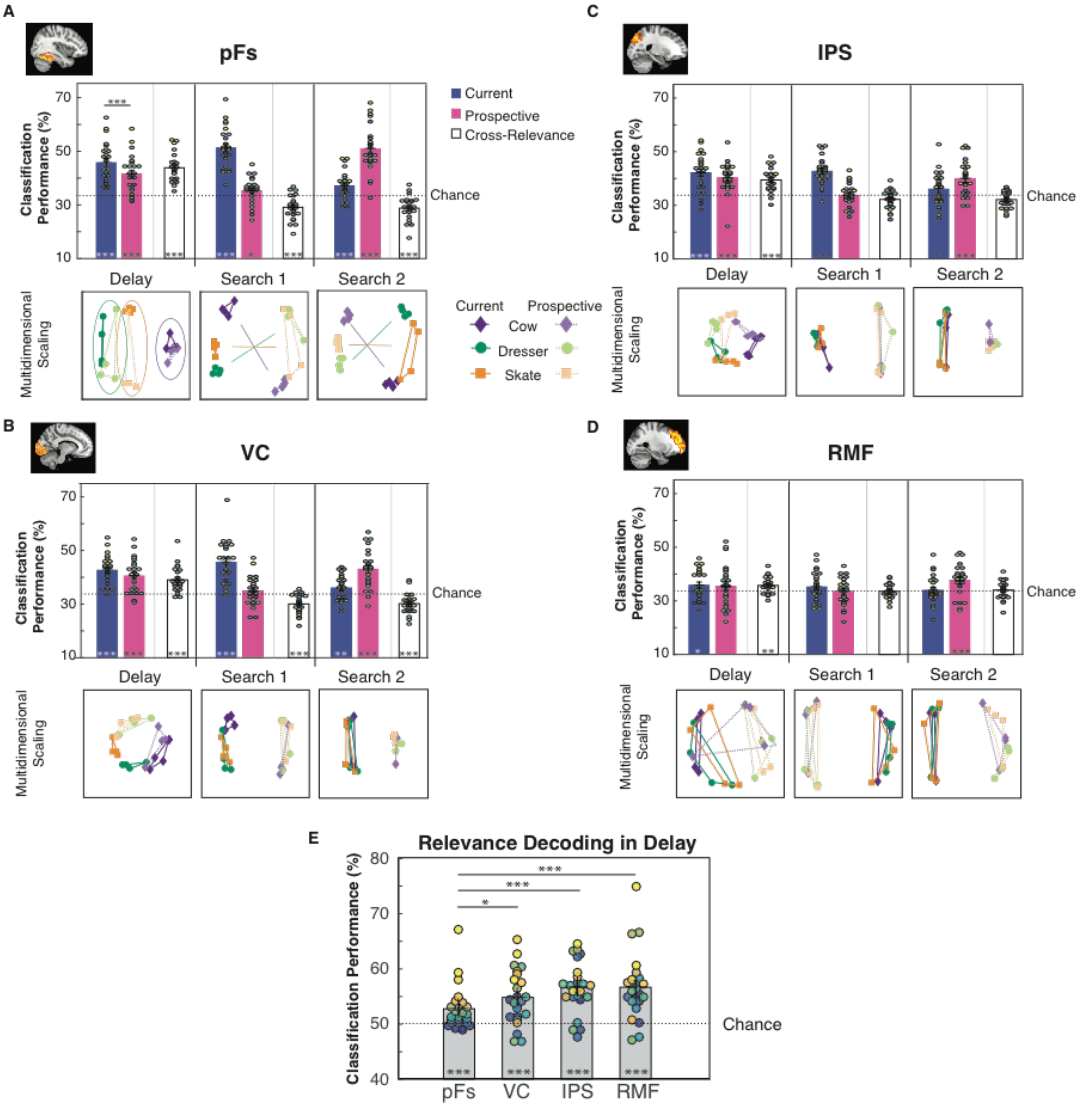
Other regions of interest

To test if our results were specific to object-selective cortex, we ran the same analyses on rostral middle frontal cortex (RMF), the intra parietal sulcus (IPS) and early visual cortex (VC). These regions have been implicated in working memory and visual search. We used anatomical masks for the regions of interests. Cortical reconstruction was performed with Freesurfer (http://surfer.nmr.mgh.harvard.edu/) and we used Freesurfer's automatic cortical parcellation algorithm (Desikan et al., 2006). We identified in each participant a region corresponding to the rostral-middle-frontal region of both left and right hemispheres, since Lee and colleagues found nearly identical results from a functionally defined lateral prefrontal cortex in a very similar experimental design (Lee et al., 2013). For the intra parietal sulcus (IPS) we used a combined probabilistic mask of IPS0, IPS1, IPS2, IPS3, IPS4 and IPS5, for our ROI of the visual cortex (VC) we combined V1v, V1d, V2v, V2d, V3v and V3d (Wang, Mruczek, Arcaro, & Kastner, 2015). Before merging the separate masks into the combined ROIS, we first applied a threshold of 20% probability to each mask to reduce the size of the masks.

As these analyses were explorative, we did not have an independent mapper and we therefore used available standardized parcellations instead (see Methods section). The results are shown in Figure S3, for comparison we also display the decoding results for pFs

(see Figure 4A). Again, we averaged for each condition the classification performance for each of the three windows of interest (Delay, Search 1 and Search 2,) and compared the performance to chance level, as well as between the two conditions. Moreover, visual cortex and IPS clearly showed evidence for category-specific information (see Figure S3BC), as indicated by above-chance object category classification for Current (blue bars) and Prospective (pink bars) conditions during the Delay (vs. 33%; VC: *t*'s > 5.55, *p*'s < 0.001, η_p^2 = 1.13 and IPS: *t*'s > 5.96, *p*'s < 0.001, η_p^2 = 1.22), for Current condition in Search 1 (VC: $t_{(1,23)}$ = 7.578, *p* < 0.001, η_p^2 = 1.55 and IPS: $t_{(1,23)}$ = 8.91, *p* < 0.001, η_p^2 = 1.82) and during Search 2 for both Current and Prospective (VC: all *t*'s > 2.74, all *p*'s < 0.006, η_p^2 = 0.56 and IPS: all *t*'s > 2.13, all *p*'s < 0.022, η_p^2 > 0.43). Weaker, but still reliable category-based differentiation also occurred in RMF, but only during the Delay period in the Current condition ($t_{(1,23)}$ = 2.18, *p* = 0.020, η_p^2 = 0.44) and during the second search in the Prospective condition ($t_{(1,23)}$ = 3.54, *p* < 0.001, η_p^2 = 0.72, Figure S3D). Higher decoding for the Current item than the Prospective item in the Delay was only apparent in pFs ($t_{(1,23)}$ = 3.22, *p* = 0.004, η_p^2 = 0.44) and not for VC, IPS and RMF (all *t*'s < 1.867 and all *p*'s > 0.075, η_p^2 < 0.38)

Finally, above-chance cross-decoding when object classification was trained and tested across relevance conditions (see Figure S3ABCD, white bars) was possible in all regions during the Delay prior to search (all *t*'s > 3.87 and all *p*'s < 0.001, η_p^2 = 0.59, we averaged the decoding performance for the two training schemes CP and PC.). However, below-chance decoding was only apparent in pFs and visual cortex (all *t*'s < -4.04 and all *p*'s < 0.001, η_p^2 < -0.82), with no such evidence in RMF (all *t*'s < 1.00 and all *p*'s > 0.327, η_p^2 < 0.20). This suggests that the category dissimilarity between current and prospective object representations was specific to posterior, more sensory-related areas, although this may of course be a matter of power and the specific stimuli we used. The pattern was supported by the multidimensional scaling plots on the similarity values of the ROIs (Figure S3ABCD, lower panels).



Supplementary Figure S3. Classification performances in (A) posterior Fusiform cortex (pFs) (B) Visual cortex (VC) (C) Intra Parietal Sulcus (IPS) (D) Rostral Middle Frontal cortex (RMF). *Within-relevance object category decoding for current (blue bars) and prospective (pink bars) items was possible in all ROIs, but the strength of the evidence varied per areas and time into the trial, with strongest object category decoding being observed in VC and pFs. Cross-relevance decoding (averaged over PC and CP; white bars) resulted in above chance decoding during the Delay period prior to search for all ROIs, but only in VC and pFs did we observe below chance decoding during Search 1 and Search 2. The multidimensional scaling plots of the similarity values of Delay, Search 1 and Search 2, demonstrated a clear distinction between current and prospective items throughout the trial. For VC and IPS the objects were also more clearly clustered according to their category than for RMF.* ***(E)*** *Relevance decoding in the Delay for all ROIs (grey bars); we were able to decode the Relevance (current versus prospective, regardless of object category) in all ROIS in the Delay prior to search. Relevance decoding was higher for VC, IPS and RMF as compared to pFs. Colored dots in the bar graphs indicate individual participant data, *p < .05, **p < .01, ***p < .001, N=24.*

In addition, we trained a classifier on Relevance only (Current or Prospective) irrespective of object category; here chance decoding was 50%. We were able to decode the relevance of the object category (current versus prospective, see Figure S3E) during the delay period prior to search in all regions (vs. 50%; all *t*'s > 3.43 and all *p*'s < 0.001, η_p^2 = 0.70). Interestingly, relevance decoding was significantly higher in the other ROIs when directly compared to the relevance decoding in pFs (all *t*'s > 2.56 and all *p*'s < 0.02, η_p^2 = 0.52), especially for IPS and RMF (η_p^2 > 0.81). This supports the idea that parietal and prefrontal populations can be sensitive to specific memory content, but may have as their primary role the planning and overall goal maintenance, while stimulus-specific representations are maintained more strongly posteriorly in sensory areas (Bugatus et al., 2017; Hebart et al., 2018; Lee et al., 2013; Postle, 2006). Taken together, the impact of relevance (Current vs Prospective) was apparent in all ROIs throughout the trial, whereas object category information was stronger for pFs, visual cortex and IPS than for RMF.