Big-Loop Recurrence within the Hippocampal System supports Integration of Information across Episodes

Raphael Koster,[1*†] **Martin J. Chadwick,**[1*†] **Yi Chen,**[2,3*] **David Berron,**[2,3,4] **Andrea Banino,**[1] **Emrah Düzel,**[2,3,5] **Demis Hassabis,**[1,6] **Dharshan Kumaran**[1,5†]

[1]DeepMind, 5 New Street Square, London EC4A 3TW, UK

[2]Institute of Cognitive Neurology and Dementia Research, Otto-von-Guericke-University Magdeburg, 39120 Magdeburg, Germany

[3]German Center for Neurodegenerative Diseases (DZNE), Site Magdeburg, 39120 Magdeburg, Germany

[4]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, 223 62 Lund, Sweden

[5]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK

[6]Gatsby Computational Neuroscience Unit, 25 Howland Street, London W1T 4JG, UK

[*]These authors contributed equally to the work.

[†]Corresponding authors: rkoster, mjchadwick, dkumaran @google.com

Summary

Recent evidence challenges the widely held view that the hippocampus is specialized for episodic memory, by demonstrating that it also underpins the integration of information across experiences. Contemporary computational theories propose that these two contrasting functions can be accomplished by big-loop recurrence, whereby the output of the system is recirculated back into the hippocampus. We use ultra-high resolution fMRI to provide support for this hypothesis, by showing that retrieved information is presented as a new input on the superficial entorhinal cortex – driven by functional connectivity between the deep and superficial entorhinal layers. Further, the magnitude of this laminar connectivity correlated with inferential performance, demonstrating its importance for behavior. Our findings offer a novel perspective on information processing within the hippocampus, and support a unifying framework in which the hippocampus captures higher-order structure across experiences, by creating a dynamic memory space from separate episodic codes for individual experiences.

Introduction

The hippocampus is widely accepted to be critical to episodic memory, by creating orthogonalized representations that minimize interference between experiences (Marr, 1971; McNaughton and Morris, 1987; Treves and Rolls, 1992; O'Reilly and McClelland, 1994; McClelland et al., 1995; Norman and O'Reilly, 2003; Yassa and Stark, 2011; Favila et al., 2016). Emerging evidence, however, suggests that the hippocampus also plays an important role in integrating information across episodes (Kumaran, 2012; Zeithamova et al., 2012; Shohamy and Turk-Browne, 2013; Collin et al., 2015; Milivojevic et al., 2015; Schlichting et al., 2015; Banino et al., 2016; Schlichting and Preston, 2017), a function that exposes a fundamental tension with its role in pattern separation (Kumaran and McClelland, 2012) – a computation that inherently disregards the commonalities between episodes (McNaughton and Morris, 1987; O'Reilly and McClelland, 1994). Recent work, therefore, presents a substantial challenge to theoretical frameworks that view the hippocampus as being specialized for episodic memory (McClelland et al., 1995; Norman and O'Reilly, 2003), with the gradual extraction of statistical structure across experiences relying on the neocortex (McClelland et al., 1995; Kumaran et al., 2016).

Recently developed computational theories (Kumaran and McClelland, 2012; also see Schapiro et al., 2017) have posited a novel mechanism that resolves the tension between episodic memory and integration across episodes. Specifically, the hippocampal system is proposed to act as a big-loop recurrent circuit, whereby the output of the system is recirculated as a new input – rather than constituting the endpoint of hippocampal processing (McClelland et al., 1995). Notably, the use of the term big-loop recurrence (Kumaran and McClelland, 2012; Schapiro et al., 2017) – which supports the appreciation of structure *across* episodes – serves to distinguish this mechanism from the well-established internal recurrence within the CA3 region of the hippocampus (Treves and Rolls, 1992), which supports *within* episode pattern completion (Fig. S1). Anatomically, recirculation mediated by big-loop recurrence could be enabled by direct projections between the input and output layers of the entorhinal cortex – which have been documented in rats (Dolorfo and Amaral, 1998; Chrobak et al., 2000; Kloosterman et al., 2004; Van Strien et al., 2009) and monkeys (Chrobak and Amaral, 2007) – or by a wider circuit involving additional neocortical regions (Kumaran and McClelland, 2012). By this algorithmic account, the hippocampus represents related experiences (e.g. AB and

BC in the paired associate inference (PAI) task; see Fig. 1A) through pattern separated codes: big-loop recurrence effectively allows the recombination of related episodes at the point of retrieval – thereby allowing appreciation of the commonalities and higher order structure present among the set of experiences (Fig. 1B, Kumaran and McClelland, 2012; also see Schapiro et al., 2017).

As an everyday example of the dynamic interplay of distinct yet related memories mediated by big-loop recurrence, imagine meeting a work colleague (Alex), and suddenly remembering meeting them for the first time in a conference in Paris the year before. This then triggers a strong memory of a romantic trip to Paris with a partner (Sam) many years previously, which itself triggers another memory of a holiday spent together in Thailand. In this example, this process effects a form of "memory chaining" whereby individual elements of episodic memories can trigger the retrieval of other memories from completely different times of our life. The big-loop recurrence account proposes that the hippocampus represents related experiences (e.g. Alex-Paris and Paris-Sam in the example above) through pattern separated codes, and that recirculation of shared components across experiences as the output back into the hippocampus (i.e. Paris) allows the retrieval of other related memories (Paris-Sam; Sam-Thailand). Whilst the example above illustrates how the algorithm allows memories to be chained together, more generally it mediates a successive search through memory with subsequent retrieval conditioned on the products of previous retrieval resulting in an iterative process consisting of cycles of memory retrieval and re-querying of the hippocampus (Kumaran, 2012; Kumaran and McClelland, 2012). This provides a powerful mechanism for generalization and inference, as exemplified by the widely-studied PAI task (Fig. 1A), as it effectively allows the recombination of related episodes at the point of retrieval, thereby allowing appreciation of the commonalities and higher order structure present among the set of experiences (Kumaran and McClelland, 2012; also see Schapiro et al., 2017).

In contrast to retrieval-based big-loop recurrent models, encoding-based models (Cohen and Eichenbaum, 1993; Gluck and Myers, 1993; Eichenbaum et al., 1999; Howard et al., 2005; Shohamy and Wagner, 2008; Schlichting and Preston, 2017) implicitly de-emphasize the notion of pattern separation – a computation viewed to be critical to episodic memory (O'Reilly and McClelland, 1994; Norman and O'Reilly, 2003) – and instead suggest that it is the overlap of hippocampal neural codes for related experiences that is critical to inferential behavior (see Fig. S2 for details).

This account is supported by several studies (Collin et al., 2015; Schlichting et al., 2015; Schlichting and Preston, 2017), indicating that encoding-based mechanisms can support inference when the stimuli have been repeated several times, giving an opportunity for the integrated representations to be formed within the hippocampus (consistent with simulated results from Schapiro et al., 2017). Thus, while this encoding-based contribution has been experimentally established, the proposed big-loop mechanism has yet to be empirically validated. Therefore, the aim of the current experiment was to test the hypothesis that the hippocampal system functions as a big-loop recurrent circuit that enables the recirculation of information from the superficial (sEC) and deep (dEC) layers of the entorhinal cortex (See Fig. S1, Kumaran and McClelland, 2012; also see Schapiro et al., 2017). To achieve this we employed the paired associate inference task (Fig. 1A, Bunsey and Eichenbaum, 1996; Zeithamova et al., 2012) – a prototypical setting in which to examine the mechanisms underlying the ability to integrate information across episodes – together with ultra-high resolution multivariate fMRI techniques which allowed us to dissociate neural activity in the input and output layers of the hippocampal system: the sEC and dEC layers of the entorhinal cortex (Maass et al., 2014), respectively (see Fig. S3 and Methods).

We focused on inference test trials (i.e. AC trials) where this circuit-level mechanism makes specific predictions (Fig. 1B). Firstly, during AC test trials requiring the integration of information across episodes (e.g. AB and BC), we should see evidence of reactivation of the unseen item (i.e. B) on the input layer (sEC) due to recirculation from the output (dEC). Secondly, we should see evidence for functional connectivity between these layers, reflecting the active recirculation of information. Thirdly, the magnitude of recirculation through this pathway connecting the output layer (dEC) to the input layer (sEC) should predict individual variation in inference performance. Critically, recirculation of the output of the hippocampal system as a new input can only exist in the presence of a big-loop mechanism, and therefore the predictions outlined above contrast with those made by encoding-based models (e.g. Shohamy and Wagner, 2008; Schlichting and Preston, 2017, also see Fig. S2). Specifically, encoding-based models are consistent with reactivation in the dEC of the shared component across related memories (i.e. the B item, shared across AB and BC pairs in the PAI task) as part of pattern completion of a composite/blended representation (i.e. ABC; see Fig. S2). However, they do not predict recirculation of this information from the dEC to the sEC mediated by connectivity

between these layers, or the functional relevance of this mechanism to behavior.

Results

Task and behavioral results

We scanned 26 subjects while they performed a PAI task (Fig. 1A). During encoding, subjects viewed pairs of images. For each pair (AB), each subject later saw another related pair (BC) which shared one image in common (B). This creates stimulus triads such that A and C are indirectly associated even though they have never been presented together. Direct test trials test whether a subject remembers the correct images being presented together (hereafter 'AB/BC test trials'), thereby assessing associative memory. By contrast, indirect test trials investigate whether the subjects are able to infer the indirect association between A and C images (hereafter 'AC test trials'), which depends on the integration of information across related episodes (i.e. AB, BC). The focus of this experiment was on the neural mechanisms present during these AC test trials.

Behaviorally, subjects had the highest accuracy for AB test trials (Fig. 2A; significant difference across all conditions: $F(2,50) = 69.88$, $p < 0.0001$; mean$\pm$ standard error of the mean; 0.89 ± 0.012, significantly larger than BC (per definition the second pair observed from each triad): $t(25) = 2.41$, $p = 0.021$ and AC: $t(25) = 10.21$, $p < 0.0001$), followed by BC (0.88 ± 0.016, significantly larger than AC: $t(25) = 8.07$, $p < 0.0001$) and was lowest for AC (0.76 ± 0.019). The reaction times followed this pattern inversely with the AB test trials being the fastest (significant difference across all conditions: $F(2,50) = 317.3$, $p < 0.0001$; 2013 ± 54, significantly faster than BC: $t(25) = 6.62$, $p < 0.0001$ and AC: $t(25) = 18.66$, $p < 0.0001$), followed by BC (2127 ± 54.3, significantly faster than AC: $t(25) = 17.88$, $p < 0.0001$) and was slowest for AC (2935 ± 47.4). Note that while incorrect trials had significantly longer RTs ($F(1,25) = 29.47$, $p < 0.0001$) the slowed response for AC trials was also observed when only considering correct trials (significantly slower than AB: $t(25) = 19.83$, $p < 0.0001$; and BC: $t(25) = 18.53$, $p < 0.0001$). This pattern of results matches those observed in a previous study, and are well accounted for by the REMERGE big-loop recurrent

model (Banino et al., 2016).

AC trials reactivate category information in the entorhinal cortex

In order to investigate the functional properties of the entorhinal cortical layers, we first needed to establish the presence of detectable scene and object signals in the fMRI data. We considered the whole-brain Scenes vs. Objects (SvO) univariate contrast during AB/BC test trials (see Fig. 1) on the group level. These trials will contain a mixture of both retrieved and visually presented information, both of which likely contribute to any resultant scene and object information. In a whole-brain analysis, we found significant SvO activation spanning the entire posterior medial temporal lobe (MTL) to the occipital lobe (Fig. 2B), consistent with the literature on scene representation (Nasr et al., 2011). Further, consistent with previous studies of the medial temporal cortex (Litman et al., 2009; Schröder et al., 2015; Berron et al., 2018), we found evidence for a significant posterior-anterior gradient of scene-to-object signal (see Methods and Fig. 2C; average $r = 0.81 \pm 0.013$, sign-rank test against 0: $z = 4.46$, $p < 0.0001$, one-tailed).

We next turned to the AC test trials (see Fig. 1). Given that no scene or object information was visually presented on these trials, the presence of any category signal can only be driven by reactivation of memory content within the medial temporal cortex. We also found evidence for a posterior-anterior gradient coding for scenes versus objects (average $r = 0.183 \pm 0.082$, sign-rank test against 0: $z = 2.17$, $p = 0.015$, one-tailed) during AC test trials, thereby demonstrating that the medial temporal cortex signal reflects memory reactivation at the time of inference. We ran a further analysis specifically focusing on AC test trial reactivation within the EC. In order to do so we first selected the 100 voxels within the EC that displayed the strongest activation for Scenes over Objects (SvO) and vice versa (OvS) during the AB/BC test trials to use as two functional regions of interest. For each subject, we then extracted the average t-value of these two sets of voxels from the SvO contrast in the AC test trials. A direct comparison revealed a significant difference in relative Scene and Object reactivation between these two (SvO and OvS) sets of voxels (mean difference: 3.13 ± 0.86, sign-rank test against 0: $z = 2.98$, $p = 0.0014$, one-tailed, see Fig. S4 for single subject activation maps). This shows that the same EC voxels that discriminate scenes and objects during presentation (AB/BC

test trials) are also involved in memory reactivation (AC test trials). These results establish that the expected signals of reactivation of the linking (B) item are present within the entorhinal cortex.

Entorhinal cortical layers have separable BOLD signal

While our previous analysis provides evidence for reactivation in the EC as a whole, our key predictions depend on the ability to detect laminar-specific signals in the deep and superficial EC layers. Detecting separate signals in different cortical layers can be challenging due to potential correlation in the BOLD signal driven by the vasculature of laminar cortex (Turner, 2002; Lawrence et al., 2017; Huber et al., 2017b; Stephan et al., 2017, also see Discussion). In order to investigate the structure of the laminar signal in our dataset, and to investigate the presence of layer-specific signals, we took a subject-specific, task-independent approach. We used an automated segmentation method (see Methods and Fig. S6) to segment the entorhinal cortex of each subject into the superficial (input) and deep (output) layers, approximately corresponding to cortical layers II/III and IV/V respectively (Maass et al., 2014). Segmentations were based on each subject's native-space high-resolution T2 images, therefore creating a set of subject-specific laminar segmentations (Fig. 3 and S3). For each voxel in both the superficial and deep entorhinal cortex, we extracted the event-specific signal for each trial across the entire experiment (including encoding trials, AB/BC, and AC test trials), thereby forming a temporal response profile. If there is detectable layer-specific signal, it should be possible to determine whether a voxel belongs to the superficial or deep layer based on its temporal responses. To directly test this question, we trained a multivariate classifier to differentiate superficial and deep voxels based on their temporal response profiles. The results were robustly significant (mean standard deviation: $85.9 \pm 0.7\%$; sign-rank test on z-scores obtained by permutation: $z = 4.46$ $p < 0.001$), demonstrating separable signal over the course of the experiment (Fig. 3B and Fig. S3B). To determine whether this was still the case when specifically focusing on the set of trials of interest to our main hypothesis (AC test trials), we repeated the analysis based only on the temporal responses derived from the set of AC test trials. The classification result still remained robust ($67.3 \pm 0.6\%$; $z = 4.46$, $p < 0.001$), validating that the signal present in each layer is separable during the set of AC test trials. Importantly, this difference cannot be attributed to mean differences across layers, but can only be based on the temporal responses, as the voxel time

courses were standardized prior to classification. We additionally directly compared the temporal signal and contrast to noise ratios (tSNR and CNR) across the layers (Welvaert and Rosseel, 2013, also see Methods and Fig. S6). We found no significant difference in CNR (based on picture presentations vs. fixations: mean±standard deviation; sEC: 0.054 ± 0.12; dEC: 0.064 ± 0.084; sign-rank test on difference: $z = 0.57$, $p = 0.57$, two-tailed), though there was a significantly stronger tSNR in the deep compared to superficial EC (mean±standard deviation; sEC: 7.32 ± 0.73; dEC: 8.41 ± 0.92; sign-rank test on difference: $z = 4.46$, $p < 0.0001$, two-tailed). This latter is consistent with some degree of MTL signal dropout, which would influence the layers closer to the cortical surface (i.e. the superficial layers, Olman et al., 2009). Importantly, however, the successful layer classification described above demonstrates that robust differences between the EC layers are still present despite this dropout. Further, it also provides additional validation of the automated laminar segmentation protocol, as separable signal could only be present if the layers have been accurately segmented.

Reactivated category information is present in the superficial entorhinal cortex at the time of inference

Having found evidence for distinct laminar signals in the entorhinal cortex, we next turned to our first main hypothesis. If episodic information is recirculated from the entorhinal output (deep) layers back to the input (superficial) layers during AC test trials, then category information (scene or object) about the linking B item should be present not only in the deep (dEC), but also the superficial (sEC) layer activations (Fig. 1B). To investigate this neural "reactivation" during AC test trials, we used multivariate voxel decoding to classify the stimulus category (object or scene) of the non-presented linking item (image B). As images A and C were always faces, successful classification can only be due to reactivation of item B. All decoding analyses were based on anatomical masks specific to each subject's anatomy (Fig. 3 and S3) in native space, and all reported results are based on unsmoothed, bilateral data (results for each hemisphere separately are reported in Table S1 and S2). First, consistent with previous multivariate studies involving the cued recall of objects and scenes (Liang and Preston, 2017), we found significant reactivation throughout the medial temporal lobe regions (MTL; Fig. 3C; for individual hippocampal subfields, see Fig. S7). Crucially, however,

we also detected significant reactivation within the sEC ($1.7\% \pm 0.07$, $z = 2.45$, $p = 0.007$, one-tailed; Fig. 3D), suggesting that information may be recirculated from the dEC back into the hippocampus via the sEC. As detailed earlier, there are potential hemodynamic confounds between the cortical layers driven by the BOLD sensitivity to the cortical vasculature (Turner, 2002; Lawrence et al., 2017; Stephan et al., 2017). In order to ensure that the classification result in the sEC was not driven by BOLD originating solely from dEC activations, we reran the classification analysis after first removing the local dEC signal from each sEC voxel (See Methods and Fig. S5). This analysis ensures that remaining signal in the sEC cannot be driven by any local spatial correlations caused by draining vein artefacts (for a conceptually similar approach, see Kok et al., 2016). The classification result in the sEC was still significant after regressing out the local dEC signal ($1.8\% \pm 0.6$, $z = 2.73$, $p = 0.0032$, one-tailed), thereby providing evidence that there is unique reactivation information present in the sEC that is unlikely to be attributed to local vasculature confounds such as draining veins (Turner, 2002; Lawrence et al., 2017; Stephan et al., 2017).

MTL informational connectivity analyses: evidence for a functional connection between the entorhinal layers

We next tested for the existence of the predicted functional connection between the entorhinal layers using a method known as informational connectivity (IC). The method has similarity to well-established functional connectivity analysis, but allows inference based on multivoxel information rather than global BOLD signal (cf. Coutanche and Thompson-Schill, 2013; Aly and Turk-Browne, 2016). This approach assesses the covariation in trial-by-trial information (decoding accuracy) between a pair of regions (Fig. 4B). If the two regions covary positively, this indicates that the regions are functionally connected, with information passing between the two (Fig. 4C). IC has previously been shown to accurately capture regional covariation in information content above and beyond univariate methods (Coutanche and Thompson-Schill, 2013; Aly and Turk-Browne, 2016; Huffman and Stark, 2017), and here we applied it to assess the flow of information during inference. We measured IC between the sEC and dEC, while partialling out the information contained in all other MTL regions, ensuring that any resulting connectivity was selective to this laminar connection, and was not mediated by any other MTL region. The laminar IC was significant (*mean*

rho $= 0.11 \pm 0.01$, $z = 4.46$, $p < 0.0001$), indicating a selective functional connection between the EC layers.

To validate that this laminar IC measure reflects a meaningful functional connection between the EC layers, we leveraged the well-established anatomical connectivity of the MTL using what we refer to as an "MTL IC direct/indirect analysis". If IC is detecting meaningful functional connections, we should find that IC is stronger between regions that are directly anatomically connected than between regions that are only indirectly connected (i.e. are only connected via other regions, as part of a larger circuit). To accomplish this, we calculated the IC between every pair of regions in the MTL (see Methods). We then separated the IC measures into pairs of regions that are known to be directly anatomically connected, and those that are only indirectly connected (Amaral and Lavenex, 2009; Libby et al., 2012; Clark and Squire, 2013, see Fig. 4D), leaving aside the laminar connection. These two sets of ICs were then averaged per subject to provide a summary of the 'direct' and 'indirect MTL' connectivity strength. Both showed a significant positive IC strength (direct: *mean rho* $= 0.096 \pm 0.002$, $z = 4.46$, $p < 0.0001$; indirect: *mean rho* $= 0.065 \pm 0.003$, $z = 4.46$, $p < 0.0001$, one-tailed), but crucially, the direct connectivity was significantly stronger than indirect connectivity ($z = 4.07$, $p < 0.0001$, one-tailed), showing that the strength of IC reflects the underlying functional connectivity of the region (see Fig. 4E). This comparison therefore allowed us to investigate whether the selective IC we find between the entorhinal layers, is more similar to a 'direct' MTL connection than an 'indirect' correction. The laminar IC was significantly higher than the indirect MTL IC ($z = 2.96$, $p = 0.0015$, one-tailed; see Fig. 4), distinguishing it from the baseline connectivity in the MTL. We further explored this connection by using a classification approach, which tested whether the laminar connection was more likely to reflect a 'direct' or 'indirect' anatomical connection. A classifier was trained to distinguish 'direct' and 'indirect' MTL connections using each subject's average IC strength as feature. This logistic ridge-regression was then applied to the laminar IC values for each subject, determining whether the laminar connection is classified as direct or indirect MTL connection. The resulting value of the regression (classifier evidence) was 0.62 (ranging from 0 ('indirect MTL') to 1 ('direct MTL'); $p < 0.024$ based on a one-tailed permutation test). Put together, these results provide empirical evidence for selective connectivity between the sEC and dEC consistent with a genuine functional connection between the layers – in keeping with the known anatomical connections between the layers (Van Strien et al., 2009).

We ran two additional analyses to ensure that this connectivity result could not be attributed to spurious local laminar correlations driven by the cortical vasculature. The first control analysis used the same procedure outlined earlier (i.e. in relation to the multivariate classification analysis), whereby the local BOLD signal from the neighbouring dEC layer was removed from the sEC on a voxel-wise basis and vice versa (see Methods). Laminar IC was recalculated using this processed data, and was still found to be significantly stronger than the indirect MTL connectivity ($z = 2.37$, $p = 0.009$, one-sided; Fig. S5). The second control analysis parcellated the EC layers each into two distinct sections (medial and lateral, see Fig. S5), and IC was recalculated between the layers across these distinct sections. There are known to be long-range anatomical connections between the layers, such that information could still in theory pass between these medial and lateral sections (Dolorfo and Amaral, 1998), but in this case there can be no contribution from any local vascular artefacts. Laminar IC across the medial and lateral sections (i.e. medial sEC with lateral dEC, and lateral sEC with medial dEC) was significantly stronger than the indirect MTL connections ($z = 2.07$, $p = 0.019$, one-sided, Fig. S5). These two control analyses provide evidence that there is a genuine selective functional connection between the EC layers (i.e. laminar connectivity) that cannot be accounted for by local hemodynamic contributions such as draining vein artefacts.

Individual variation in laminar connectivity predicts strength of input back into the hippocampus

Having established that the sEC contains reactivated information at the time of inference, and that there is a selective functional connectivity between the entorhinal layers, we next tested whether the sEC reactivation is driven by information flowing through this 'big-loop' pathway, in line with our hypothesis (Fig. 1). We leveraged the individual variation in reactivation and IC strength in order to address this question. If information flows from the dEC to the sEC, then we would expect to see a positive correlation between laminar IC strength and sEC reactivation strength. Further, big-loop recirculation predicts that the correlation between the magnitude of laminar IC and reactivation strength in the input layer (sEC) should be significantly stronger than with the output layer (dEC). This is because the strength of sEC reactivation should depend specifically on big-loop recurrence indexed through laminar IC, whereas the strength

of dEC reactivation likely reflects the more conventional mechanism of simple associative retrieval – where a subject retrieves the individual associate pairs AB or BC from the A and C items presented on the screen during AC test trials. Fig. S8 displays simulation results confirming that, when individual differences in these two processes are independent, we should find differential correlations between laminar IC and sEC vs. dEC reactivation.

As predicted, we found a significant positive correlation between the sEC reactivation strength and laminar IC (Fig. 5, $r = 0.406$, $p = 0.018$; one-tailed permutation tests used for all across-subject correlations). By contrast, the dEC was not significantly correlated with laminar IC ($r = -0.202$, $p = 0.838$), and a direct comparison of the two revealed that the sEC was significantly more highly correlated with laminar IC than the dEC (difference in $r = 0.607$, $p = 0.014$). This provides evidence consistent with the hypothesis that selective laminar connectivity is indeed driving the sEC reactivation. Note that this difference can not be explained by a generic difference in signal intensity between the layers. There is no significant difference in CNR between the two layers, and tSNR is stronger in the dEC than in the sEC, which would predict a bias in sensitivity in the opposite direction to our results (see Fig. S6 for details). To ensure that these correlations results were not influenced by the positive correlation between the sEC and dEC reactivation strengths ($r = 0.199$, $p = 0.168$), we ran a further partial correlation analysis where for each correlation we controlled for reactivation strength in the other layer. This analysis revealed the same pattern of result (sEC: $r = 0.464$, $p = 0.009$; dEC: $r = -0.315$, $p = 0.939$; difference in $r = 0.779$, $p = 0.003$), demonstrating that the relationship between laminar IC strength and sEC reactivation cannot be explained by confounding correlations with the dEC. Notably, the finding of a clear dissociation between the entorhinal layers result provides additional strong evidence against the data being driven by vasculature confounds, as an effect that is uniquely attributable to the sEC cannot be driven by spurious BOLD confounds deriving from the dEC.

One further prediction of the big-loop hypothesis is that information recirculated from the dEC to the sEC should then propagate back into the hippocampus – specifically into subregions viewed to instantiate pattern separated representations for individual experiences ('episode space'; see Fig. S2), namely the dentate gyrus (DG) and CA3 (see Fig. S1) – thereby allowing retrieved content to requery related experiences stored in these regions (Kumaran and McClelland, 2012). We tested this prediction using the same individual variation data, by correlating the laminar IC strength with

the average IC strength between the sEC and the DG&CA3. This revealed a significant positive correlation ($r = 0.391$, $p = 0.026$), further supporting the hypothesis that the big-loop pathway drives recirculation of information back into the hippocampus. As each of these connections shares the sEC in common, they are not fully independent, which could artificially inflate the correlation. We therefore assessed a baseline control correlation between laminar IC and the average IC strength between the sEC and CA1&SUB. While these hippocampal subfields do share an anatomical connection with the sEC, these connections should not form part of the big-loop recurrent pathway (Kumaran, 2012; Kumaran and McClelland, 2012). Thus, if the sEC to DG&CA3 IC reflects the predicted operation of the big-loop pathway, the strength of this connection should be significantly greater than IC between the sEC and CA1&SUB. Notably, these latter connections share the same statistical dependence on the sEC, and therefore provide a stringent baseline. Further, this comparison provides a strong test of the predicted pathway from the sEC back into the hippocampus. Laminar IC did not significantly correlate with the sEC/CA1&SUB IC strength ($r = -0.214$, $p = 0.82$), and crucially, it was significantly weaker than the correlation between laminar IC and the sEC/DG&CA3 pathway (difference in $r = 0.605$, $p = 0.016$). These results clearly support the hypothesis that information is recirculated from the dEC to the sEC back into the hippocampal subfields CA3 and DG during performance of AC inference test trials.

The results presented above point towards the conclusion that the laminar connectivity identified reflects a direct functional pathway between the dEC and sEC – rather than a larger pathway mediated by MTL regions such as the PHC or PRC. To examine this further in the current context of recirculation of information back into the hippocampus, we also considered an alternative model of information flow incorporating a longer cortical loop through the perirhinal and parahippocampal cortex (PRC; PHC). Contrary to the predictions of this alternative model, we did not find a significant correlation between sEC reactivation and sEC/PRC IC ($r = -0.083$, $p = 0.69$) or sEC/PHC IC ($r = 0.06$, $p = 0.776$), nor was there a significant correlation between sEC/DG&CA3 and sEC/PRC ($r = 0.207$, $p = 0.306$) or sEC/PHC ($r = -0.078$, $p = 0.713$) IC. Thus, these results do not provide support for any longer cortical pathway, and are instead consistent with information flowing directly from the dEC to the sEC. Further, we ran each analysis outlined above, additionally partialling out both the sEC/PRC and sEC/PHC IC. Even with these extra controls, each analysis was still significant (Table S3) demonstrating that they are not influenced by any extraneous signals arising in PRC or PHC.

Recirculation predicts individual variation in inference performance

Having established the existence of the predicted functional pathway, we next turned to the final prediction of our model – that big-loop recirculation should predict behavioral inference performance. To test this prediction, we again leveraged the individual variation in our data, and correlated the strength of laminar IC with average performance on the AC test trials (Fig. 5C). These variables were significantly positively correlated ($r = 0.378$, $p = 0.028$), indicating that the degree of recirculation has an impact on inference ability. By contrast, dEC reactivation – which we use as a proxy for the strength of conventional associative retrieval of the AB and BC pairs through pattern completion elicited by the presence of A and C items on the screen during AC test trials – did not significantly correlate with behavioral AC inference performance ($r = -0.197$, $p = 0.833$). Indeed, a direct comparison revealed that the correlation with AC test performance was significantly higher with laminar IC compared to dEC reactivation (difference in $r = 0.574$, $p = 0.021$). Additional control analyses demonstrated that these results could not be accounted for by any confounding correlations between the laminar IC and dEC reactivation ($r = -0.133$, $p = 0.736$; Table S4), or by additional

pathways running through the PHC/PRC (Table S4). To further assess the laminar-specific contributions to behavior, we correlated AC inference performance with both sEC and dEC reactivation strength. If AC inference performance is selectively driven by big-loop recurrence passing information to the sEC, we would expect to see a significantly higher behavioral correlation with sEC than dEC reactivation. This was the case both with (difference in $r = 0.606$, $p = 0.017$) and without (difference in $r = 0.481$, $p = 0.045$) partialling out the reactivation strength in the opposing layer. This difference was driven by a marginally positive relationship between AC inference performance and sEC reactivation (while partialling out dEC: $r = 0.337$, $p = 0.05$; without partialling out dEC: $r = 0.284$, $p = 0.079$). By contrast, the dEC was not positively correlated with AC inference performance either with ($r = -0.27$, $p = 0.9$) or without ($r = -0.197$, $p = 0.831$) partialling. This pattern of results was still present even when additionally partialling out the IC from PRC and PHC (AC performance and sEC reactivation: $r = 0.337$, $p = 0.058$; AC performance with dEC reactivation: $r = -0.27$, $p = 0.895$; difference: $r = 0.607$, $p = 0.022$). These results suggest that inference is dissociably related to information in the entorhinal input layers. Put together, these results reveal a direct correspondence between the strength of big-loop recirculation and overall inference performance across subjects, which could not be accounted for by variations in general retrieval strength. This is striking evidence in support of the unique functional role of the big-loop mechanism in episodic inference.

Discussion

We set out to provide evidence for the existence of a big-loop recurrent circuit using ultra-high resolution fMRI and multivoxel pattern analysis, applied to the paired associative inference (PAI) task. Solving the inference trials of this task requires the integration of information across temporally distinct episodes. If such integration requires big-loop recurrence, we should see evidence of recirculation of information from the output (dEC) back to the input (sEC). A series of targeted analyses revealed a coherent set of results, each consistent with the presence of a big-loop mechanism. First, we found evidence for reactivation of the unseen 'bridging' element (e.g. B item in AC test trial) required for successful inference not only in the region that receives the output of the hippocampus (dEC), but also

in the layer processing the input (sEC). This reactivation is only expected if a recirculation mechanism is recruited. Second, we found that the layers of the entorhinal cortex displayed informational connectivity (IC) consistent with a selective connection between the two. Further, an analysis of the individual variation in layer reactivation and IC strength provided evidence for a flow of information, from the dEC back into the sEC, consistent with the operation of a big-loop pathway. Third, laminar IC predicted the flow of information from the input layer (sEC) back into the dentate gyrus and CA3, consistent with the hypothesis that big-loop recurrence allows retrieved content to requery related experiences stored in these regions. Finally, the strength of information flow through the entorhinal big-loop predicted individual variation in inference performance. Put together, our findings offer the first evidence in any species of the functional importance of big-loop recurrence – involving recirculation of the output of the hippocampal system from the deep layer of the entorhinal cortex as a new input on the superficial layer – and establishes its link to successful performance in a classic task requiring inferential behavior. Importantly, the demonstration of big-loop recurrence bolsters theoretical frameworks that propose that the hippocampus is able to support two functions that place seemingly opposing demands on the system – episodic memory and the rapid integration of information across episodes (Kumaran and McClelland, 2012; also see Schapiro et al., 2017). Whilst these accounts have received support from computational modelling of behavioral and neural data (Kumaran and McClelland, 2012; Banino et al., 2016), they have lacked the critical empirical evidence that we provide here concerning the core underlying circuit-level mechanism, namely big-loop recurrence.

The set of results critically depends on the ability to meaningfully differentiate the signal present within the superficial and deep layers of the entorhinal cortex, as well as to detect a meaningful covariance structure between the information contained within the layers. It is therefore crucial to carefully consider the known problems of laminar fMRI introduced by the potential for local vasculature confounds. Specifically, BOLD is sensitive to signals from draining veins that run perpendicular to the cortical layers, towards the pial surface, which can lead to spurious correlations between the layers (Turner, 2002; Lawrence et al., 2017; Huber et al., 2017a; Stephan et al., 2017), making it potentially difficult to detect layer-specific signals or laminar connectivity. While alternative methods exist which provide superior specificity, they come at the cost of lower sensitivity and acquisition speed (Huber et al., 2017a,b), which can be prob-

lematic in regions sensitive to signal dropout, such as the MTL. However, despite the lower specificity of BOLD, a growing number of studies have found robust laminar-specific activation or information using BOLD imaging (Olman et al., 2012; Maass et al., 2014; De Martino et al., 2015; Muckli et al., 2015; Fracasso et al., 2016; Kok et al., 2016). Recent work, therefore, provides evidence that laminar-specific signals can be detected using BOLD fMRI.

To ensure this was true of our own data, we used a task-independent analysis to determine the presence of laminar-specific signals within our data. Specifically, a classifier trained on the set of voxel temporal response profiles, was able to differentiate superficial from deep voxels (see Fig. 3B and S3B), clearly demonstrating that the entorhinal layers contain separable BOLD signal. Further, for each analysis, we took additional steps to ensure that the results could not be attributed to vasculature-driven laminar confounds. First, we demonstrated that sEC reactivation was still present even after removing the spatially proximal dEC signal from every sEC voxel (for a conceptually similar approach, see Kok et al., 2016). This ensured that the sEC signals could not be driven by hemodynamic signals originating in neighbouring dEC voxels, as might occur in the case of draining vein artefacts (Kok et al., 2016; Huber et al., 2017a,b; Lawrence et al., 2017). Second, we applied the same local-signal correction prior to measuring the IC between the sEC and dEC, and still found evidence for a significant selective laminar connection. We also applied a second control analysis to rule out local vasculature influences on laminar connectivity, by explicitly measuring only non-local connectivity between the entorhinal layers. This was achieved by partitioning each layer into medial and lateral portions, and assessing laminar IC across different portions. Despite the fact that only non-local connectivity could be detected using this method, we still found a significant effect – consistent with the known existence of non-local anatomical connections (Dolorfo and Amaral, 1998). Each of these control analyses removes the potential contribution of local vasculature confounds, and together they provide clear evidence for the presence of a genuine functional connection between the superficial and deep EC layers. In addition to these explicit control analyses, there are also some key results that cannot be explained by vasculature confounds. First, we found that individual variation in the strength of laminar IC correlates selectively with sEC, and not dEC reactivation. If the sEC signal and laminar IC were both influenced by hemodynamic artefacts originating in the dEC, it would not be possible to find any such laminar-specific correlations, as the two layers would covary in their reactivation strength. Second, laminar

IC correlated with AC performance, while dEC reactivation did not. Again, if the sEC reactivation and connectivity signals were spuriously influenced by signals from the dEC, this selective correlation with performance would not be present. Put together, these stringent control analyses, and laminar-specific correlation results converge in providing strong evidence for the predicted recirculation signals that cannot be accounted for by any signals originating in the deep entorhinal layers.

We consider two further potential limitations of our data. First, when dealing with sub-millimeter resolution imaging and laminar separation, results are more prone to noise effects from imperfect motion correction or coregistration. In order to apply stringent quality control, all masks were manually inspected and corrected when necessary prior to analysis. To highlight the quality of the masks we display single subject level masks for every subject in EPI space in the supplemental materials (Fig. S3). Notably, successful differentiation of the BOLD signal depends on an accurate segmentation of the layers, as well as coregistration with the EPI data. If any of these steps resulted in substantive misalignment, the detection of layer-specific BOLD would not be possible. Further, any serious residual head motion artefacts remaining after preprocessing would also reduce our ability to detect layer-specific BOLD. As described in the previous passage, we found that a classifier could successfully differentiate the BOLD signal in the two EC layers, which provides a robust quality control check to rule out any significant impact from preprocessing error or residual motion. Second, while the absolute classification accuracies in each region are low, they are comparable with other decoding analyses within the region (e.g. Chadwick et al., 2010; Bonnici et al., 2012). Further, there is no straightforward relationship between classification accuracy and the underlying effect size or biological interpretation (Hebart and Baker, 2017). Indeed, the overall classification accuracy is not typically the measure of interest, as it is not itself a statistical test, and cannot be interpreted as a measure of effect size (Pereira et al., 2009). Rather, it is the nature of the statistical significance test performed on the classification accuracy that is key for appropriate interpretation, which in our case is at the group level. It is worth noting that, despite the low average decoding accuracy within the sEC, the variance in decoding across individuals correlated with laminar connectivity and inference, in each case with a medium effect size (i.e. between 0.3 and 0.5). This indicates that the variance across a relatively narrow band of decoding accuracies is nevertheless meaningful with respect to other neural metrics and behavior.

We identified selective functional connectivity – referred to as laminar connectivity – between the dEC and sEC through an IC analysis focused on the MTL. Results from a specifically designed MTL IC direct/indirect analysis were consistent with this laminar connectivity reflecting a direct connection between the dEC and sEC, rather than a longer pathway involving the PHC or PRC. Since the major neocortical inputs and outputs to/from the ERC pass through the PRC and PHC (Libby et al., 2012), controlling for information contained within these two regions in the MTL IC analysis – and additionally in an analysis linking the strength of sEC reactivation to laminar connectivity strength – also suggests that the laminar connection cannot be explained by additional regions that project through the PRC/PHC. Taken together, a parsimonious interpretation of these results is that the laminar connectivity identified reflects a direct functional pathway between the dEC and sEC – rather than a longer pathway mediated by an additional neocortical region to which the ERC projects. However, there are two more subtle alternatives that our analyses cannot conclusively rule out. First, laminar connectivity could be mediated by a part (i.e subsection) of an existing MTL region that is not adequately captured by our regions of interest. Second, the laminar connectivity identified could in theory be mediated by a region outside the MTL, that the ERC projects to via other neocortical or subcortical regions other than the PRC or PHC. Nevertheless, it is worth noting that the logic of our MTL IC direct/indirect analysis – specifically, that regions within the MTL that are directly connected have significantly greater connection strengths than those that are connected indirectly via an additional MTL region – may well hold true for brain structures outside the MTL regions our analysis was focused on. Notably, however, this argument relies on an assumption, albeit a plausible one: that the connectivity-related properties of these extra-MTL structures are broadly similar to the MTL regions examined – for example, the connection strengths to a region putatively mediating laminar connectivity are not substantially higher than the connections within the MTL. Hence, given the inherent limitations of fMRI connectivity analyses alluded to above, we cannot definitely rule out the possibility that the laminar connectivity observed does indeed involve additional regions (i.e. is indirect) – rather than being mediated by direct projections from the dEC to SEC. Importantly, the big-loop theory as originally proposed viewed the critical algorithmic component of recirculation between the hippocampal system output layer (dEC) and input layer (SEC) as being instantiated either via direct anatomical connections within the ERC known to be present (Van Strien et al., 2009; Dolorfo and Amaral, 1998; Kloosterman et al., 2004; Chrobak et al., 2000, 2007), or via additional neocortical regions (see Kumaran and

McClelland, 2012).

Whilst our results provide support that big-loop recurrence at the point of retrieval (i.e. during performance of AC test trials) supports performance in the PAI task, it is important to note that empirical evidence suggests that encoding-based mechanisms may also play a role (Zeithamova and Preston, 2010; Kumaran, 2012; Kumaran and McClelland, 2012; Zeithamova et al., 2012; Collin et al., 2015; Milivojevic et al., 2015; Schlichting et al., 2015; Banino et al., 2016; Schlichting and Preston, 2017). Indeed, whilst our results are consistent with two computational models that incorporate big-loop recurrence (Kumaran and McClelland, 2012; Schapiro et al., 2017), these models have a key difference: the latter model additionally incorporates a mechanism that allows the gradual encoding of overlapping representations that stably capture the statistics and commonalities among related experiences (Eichenbaum et al., 1999; Shohamy and Turk-Browne, 2013; Schapiro et al., 2014, 2017). This is mediated by relatively slow learning within a pathway directly connecting the EC to the CA1 region of the hippocampus (termed, the monosynaptic pathway; see Fig. S1). As such, the model by Schapiro et al. (2017) can be considered a hybrid model that incorporates the principles of big-loop recurrence (Kumaran and McClelland, 2012) as well as an encoding-based mechanism. Given the slow learning nature of the encoding based mechanism, big-loop recurrence is viewed to be critical to rapidly appreciating the structure among experiences in the PAI and related tasks after few stimulus repetitions before the MSP has had a chance to learn – that is, after a single exposure in computational simulations (Schapiro et al., 2017), and 2 exposures in our experiment. After more extended stimulus repetitions, the MSP is viewed to play a dominant role in capturing and exploiting the structure among related experiences, though big-loop recurrence is still viewed to contribute (Schapiro et al., 2017). Future work should aim to distinguish between these two big-loop recurrence models by isolating the specific operation of the MSP pathway. More generally, it will be important to characterize the conditions that favour capturing the structure of experiences on the fly through big-loop recurrent dynamics, as opposed to representing them through static neural codes – considering factors such as task parameters (e.g. the number of item repetitions), the rate of change of the environment (Gershman et al., 2017), and differences along the anterior-posterior axis of the hippocampus (Collin et al., 2015; Schlichting et al., 2016).

Recent computational models have been proposed that marry the competing functions of the hippocampus in episodic

memory and the integration of information across episodes (Kumaran and McClelland, 2012; Schapiro et al., 2017). Critically, these have incorporated big-loop recurrence as a key component, which until now has lacked empirical support. In providing the first empirical evidence for big-loop recurrence, our study encourages a rethinking of classical theories of information flow within the hippocampal system. Further, our work provides a novel algorithmic framework whereby the hippocampus can rapidly represent a set of related experiences through pattern separated codes – thereby preserving memory for individual experiences – while big-loop recurrence creates a dynamic memory space at the point of retrieval, thereby allowing the rapid appreciation of their commonalities and higher order structure. Interestingly, the algorithm realized through big-loop recurrence (Kumaran and McClelland, 2012) has striking parallels with cutting edge machine learning neural network architectures that utilize external memory to solve problems of real world relevance, typically referred to as "question answering" – where a set of sentences must be retained over a period time following which questions are posed (see Fig. S9). Notably, these tasks pose exactly the same challenge that we highlight here (e.g. Sukhbaatar et al., 2015, see also Graves et al., 2016) – how to integrate information across separate episodes whilst preserving their individual features – and would be predicted to be difficult for encoding-based models (e.g. Shohamy and Wagner, 2008; Schlichting and Preston, 2017) to solve (see Fig. S2). In the future, it will be important to establish the relative contribution of big-loop recurrence, as compared to encoding-based mechanisms, to combining information across multiple related experiences in the service of adaptive behaviour.

Author Contributions

D.K., R.K., M.J.C., D.H., Y.C., A.B. conceived the project. R.K. and Y.C. acquired the data. R.K., M.J.C, Y.C., D.B., D.K., A.B. analyzed and interpreted the data. D.K., D.H., E.D. supervised the project. All authors contributed to writing the manuscript.

Acknowledgements

We thank the Leibniz Institute for Neurobiology (LIN) for access to the 7T scanner and Caswell Barry, Matthew Botvinick, Zeb Kurth-Nelson, Adam Santoro, Christopher Summerfield and Jane X. Wang for helpful discussion. ED was supported by the German Research foundation (SFB 779 A07).

Declaration of Interest

The authors declare no conflict of interest.

References

Aly M, Turk-Browne NB (2016) Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences* 113:E420–E429.

Amaral D, Lavenex P (2009) *Hippocampal Neuroanatomy* Oxford University Press.

Avants B, Epstein C, Grossman M, Gee J (2008) Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* 12:26–41.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage* 54:2033–2044.

Banino A, Koster R, Hassabis D, Kumaran D (2016) Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports* 6.

Berron D, Vieweg P, Hochkeppler A, Pluta J, Ding SL, Maass A, Luther A, Xie L, Das S, Wolk D et al. (2017) A protocol for manual segmentation of medial temporal lobe subregions in 7tesla mri. *NeuroImage: Clinical* .

Berron D, Neumann K, Maass A, Schütze H, Fliessbach K, Kiven V, Jessen F, Sauvage M, Kumaran D, Düzel E (2018) Age-related functional changes in domain-specific medial temporal lobe pathways. *Neurobiology of Aging* .

Bonnici HB, Chadwick M, Kumaran D, Hassabis D, Weiskopf N, Maguire EA (2012) Multi-voxel pattern analysis in human hippocampal subfields. *Frontiers in human neuroscience* 6:290.

Brodeur MB, Dionne-Dostie E, Montreuil T, Lepage M (2010) The bank of standardized stimuli (boss), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one* 5:e10773.

Bunsey M, Eichenbaum H (1996) Conservation of hippocampal memory function in rats and humans. *Nature* 379:255.

Burton AM, White D, McNeill A (2010) The glasgow face matching test. *Behavior Research Methods* 42:286–291.

Chadwick MJ, Hassabis D, Weiskopf N, Maguire EA (2010) Decoding individual episodic memory traces in the human hippocampus. *Current Biology* 20:544–547.

Chrobak JJ, Amaral DG (2007) Entorhinal cortex of the monkey: Vii. intrinsic connections. *Journal of comparative neurology* 500:612–633.

Chrobak JJ, Lörincz A, Buzsáki G (2000) Physiological patterns in the hippocampo-entorhinal cortex system. *Hippocampus* 10:457–465.

Clark RE, Squire LR (2013) Similarity in form and function of the hippocampus in rodents, monkeys, and humans. *Proceedings of the National Academy of Sciences* 110:10365–10370.

Cohen NJ, Eichenbaum H (1993) *Memory, Amnesia and the Hippocampal System* MIT Press, Cambridge, MA.

Collin SH, Milivojevic B, Doeller CF (2015) Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience* 18:1562–1564.

Coutanche MN, Thompson-Schill SL (2013) Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in Human Neuroscience* 7:15.

De Martino F, Moerel M, Ugurbil K, Goebel R, Yacoub E, Formisano E (2015) Frequency preference and attention effects across cortical depths in the human primary auditory cortex. *Proceedings of the National Academy of Sciences* 112:16036–16041.

Dolorfo CL, Amaral DG (1998) Entorhinal cortex of the rat: organization of intrinsic connections. *Journal of Comparative Neurology* 398:49–82.

Eichenbaum H, Dudchenko P, Wood E, Shapiro M, Tanila H (1999) The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron* 23:209–226.

Favila SE, Chanales AJ, Kuhl BA (2016) Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications* 7.

Fracasso A, Petridou N, Dumoulin SO (2016) Systematic variation of population receptive field properties across cortical depth in human visual cortex. *Neuroimage* 139:427–438.

Gershman SJ, Monfils MH, Norman KA, Niv Y (2017) The computational nature of memory modification. *eLife* 6:e23763.

Gluck MA, Myers CE (1993) Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* 3:491–516.

Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J et al. (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538:471–476.

Hasselmo ME (2006) The role of acetylcholine in learning and memory. *Current opinion in neurobiology* 16:710–715.

Hasselmo ME, McClelland JL (1999) Neural models of memory. *Current Opinion in Neurobiology* 9:184–188.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2430.

Hebart MN, Baker CI (2017) Deconstructing multivariate decoding for the study of brain function. *NeuroImage* .

Howard MW, Fotedar MS, Datey AV, Hasselmo ME (2005) The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological Review* 112:75.

Huber L, Handwerker DA, Jangraw DC, Chen G, Hall A, Stüber C, Gonzalez-Castillo J, Ivanov D, Marrett S, Guidi M et al. (2017b) High-resolution cbv-fmri allows mapping of laminar activity and connectivity of cortical input and output in human m1. *Neuron* 96:1253–1263.

Huber L, Uludağ K, Möller HE (2017a) Non-bold contrast for laminar fmri in humans: Cbf, cbv, and cmro2. *Neuroimage* .

Huffman DJ, Stark CE (2017) The influence of low-level stimulus features on the representation of contexts, items, and their mnemonic associations. *NeuroImage* .

In MH, Speck O (2012) Highly accelerated psf-mapping for epi distortion correction with improved fidelity. *Magnetic Resonance Materials in Physics, Biology and Medicine* 25:183–192.

Ketz N, Morkonda SG, O'Reilly RC (2013) Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology* 9:e1003067.

Kloosterman F, van Haeften T, Lopes da Silva FH (2004) Two reentrant pathways in the hippocampal-entorhinal system. *Hippocampus* 14:1026–1039.

Kok P, Bains LJ, van Mourik T, Norris DG, de Lange FP (2016) Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology* 26:371–376.

Kumaran D (2012) What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience* 6:157.

Kumaran D, Hassabis D, McClelland JL (2016) What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences* 20:512–534.

Kumaran D, McClelland JL (2012) Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological Review* 119:573.

Lawrence SJ, Formisano E, Muckli L, de Lange FP (2017) Laminar fmri: applications for cognitive neuroscience. *Neuroimage* .

Liang JC, Preston AR (2017) Medial temporal lobe reinstatement of content-specific details predicts source memory. *Cortex* 91:67–78.

Libby LA, Ekstrom AD, Ragland JD, Ranganath C (2012) Differential connectivity of perirhinal and parahippocampal cortices within human hippocampal subregions revealed by high-resolution functional imaging. *Journal of Neuroscience* 32:6550–6560.

Litman L, Awipi T, Davachi L (2009) Category-specificity in the human medial temporal lobe cortex. *Hippocampus* 19:308–319.

Maass A, Berron D, Libby LA, Ranganath C, Düzel E (2015) Functional subregions of the human entorhinal cortex. *Elife* 4:e06426.

Maass A, Schütze H, Speck O, Yonelinas A, Tempelmann C, Heinze HJ, Berron D, Cardenas-Blanco A, Brodersen KH, Stephan KE et al. (2014) Laminar activity in the hippocampus and entorhinal cortex related to novelty and episodic encoding. *Nature Communications* 5:5547.

Marr D (1971) Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci* 262:23–81.

McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102:419.

McNaughton BL, Morris RG (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences* 10:408–415.

Milivojevic B, Vicente-Grabovetsky A, Doeller CF (2015) Insight reconfigures hippocampal-prefrontal memories. *Current Biology* 25:821–830.

Muckli L, De Martino F, Vizioli L, Petro LS, Smith FW, Ugurbil K, Goebel R, Yacoub E (2015) Contextual feedback to superficial layers of v1. *Current Biology* 25:2690–2695.

Mumford JA, Turner BO, Ashby FG, Poldrack RA (2012) Deconvolving {BOLD} activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* 59:2636 – 2643.

Nasr S, Liu N, Devaney KJ, Yue X, Rajimehr R, Ungerleider LG, Tootell RB (2011) Scene-selective cortical regions in human and nonhuman primates. *Journal of Neuroscience* 31:13771–13785.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping* 15:1–25.

Norman KA, O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review* 110:611–646.

Olman CA, Davachi L, Inati S (2009) Distortion and signal loss in medial temporal lobe. *PloS one* 4:e8160.

Olman CA, Harel N, Feinberg DA, He S, Zhang P, Ugurbil K, Yacoub E (2012) Layer-specific fmri reflects different neuronal computations at different depths in human v1. *PloS one* 7:e32536.

O'Reilly RC, McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* 4:661–682.

O'Reilly RC, Rudy JW (2001) Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review* 108:311.

Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE (2011) *Statistical parametric mapping: the analysis of functional brain images* Elsevier.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fmri: a tutorial overview. *Neuroimage* 45:S199–S209.

Qian J, Hastie T, Friedman J, Tibshirani R, Simon N (2013) Glmnet for matlab. *http://www.stanford.edu/∼hastie/glmnet_matlab/* .

Schapiro AC, Gregory E, Landau B, McCloskey M, Turk-Browne NB (2014) The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience* 26:1736–1747.

Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372:20160049.

Schlichting M, Preston A (2017) *The Hippocampus and Memory Integration: Building Knowledge to Navigate Future Decisions*, pp. 405–437 Springer International Publishing, Cham.

Schlichting ML, Guarino KF, Schapiro AC, Turk-Browne NB, Preston AR (2016) Hippocampal structure predicts statistical learning and associative inference abilities during development. *Journal of Cognitive Neuroscience* .

Schlichting ML, Mumford JA, Preston AR (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications* 6.

Schröder TN, Haak KV, Jimenez NIZ, Beckmann CF, Doeller CF (2015) Functional topography of the human entorhinal cortex. *Elife* 4.

Shohamy D, Turk-Browne NB (2013) Mechanisms for widespread hippocampal involvement in cognition. *Journal of Experimental Psychology: General* 142:1159.

Shohamy D, Wagner AD (2008) Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* 60:378–389.

Smith S, Jenkinson M, Woolrich M, Beckmann C, Behrens T, Heidi J, Bannister P, Luca M, Drobnjak I, Flitney D, Niazy R, Saunders J, Vickers J, Zhang Y, Stefano N, Brady J, Matthews P (2004) Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23:S208–S219.

Stephan K, Petzschner F, Kasper L, Bayer J, Wellstein K, Stefanics G, Pruessmann K, Heinzle J (2017) Laminar fmri and computational theories of brain function. *NeuroImage* .

Striem-Amit E, Ovadia-Caro S, Caramazza A, Margulies DS, Villringer A, Amedi A (2015) Functional connectivity of visual cortex in the blind follows retinotopic organization principles. *Brain* 138:1679–1695.

Sukhbaatar S, Weston J, Fergus R et al. (2015) End-to-end memory networks In *Advances in Neural Information Processing Systems*, pp. 2440–2448.

Thesen S, Heid O, Mueller E, Schad LR (2000) Prospective acquisition correction for head motion with image-based tracking for real-time fmri. *Magnetic resonance in medicine* 44:457–465.

Treves A, Rolls ET (1992) Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus* 2:189–199.

Tuğran E, Kocak M, Mirtagioğlu H, Yiğit S, Mendes M (2015) A simulation based comparison of correlation coefficients with regard to type i error rate and power. *Journal of Data Analysis and Information Processing* 3:87.

Turner R (2002) How much cortex can a vein drain? downstream dilution of activation-related cerebral blood oxygenation changes. *Neuroimage* 16:1062–1067.

Van Strien NM, Cappaert N, Witter MP (2009) The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. *Nature Reviews Neuroscience* 10:272–282.

Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017) Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* 145, Part B:166 – 179.

Welvaert M, Rosseel Y (2013) On the definition of signal-to-noise ratio and contrast-to-noise ratio for fmri data. *PloS one* 8:e77089.

Wisse L, Gerritsen L, Zwanenburg JJ, Kuijf HJ, Luijten PR, Biessels GJ, Geerlings MI (2012) Subfields of the hippocampal formation at 7t mri: In vivo volumetric assessment. *Neuroimage* 61:1043–1049.

Yassa MA, Stark CE (2011) Pattern separation in the hippocampus. *Trends in Neurosciences* 34:515–525.

Yushkevich PA, Pluta JB, Wang H, Xie L, Ding S, Gertje EC, Mancuso L, Kliot D, Das SR, Wolk DA (2015) Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping* 36:258–287.

Zeithamova D, Dominick AL, Preston AR (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* 75:168–179.

Zeithamova D, Preston AR (2010) Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience* 30:14676–14684.

Zhang D, Snyder AZ, Fox MD, Sansbury MW, Shimony JS, Raichle ME (2008) Intrinsic functional relations between human cerebral cortex and thalamus. *Journal of Neurophysiology* 100:1740–1748.

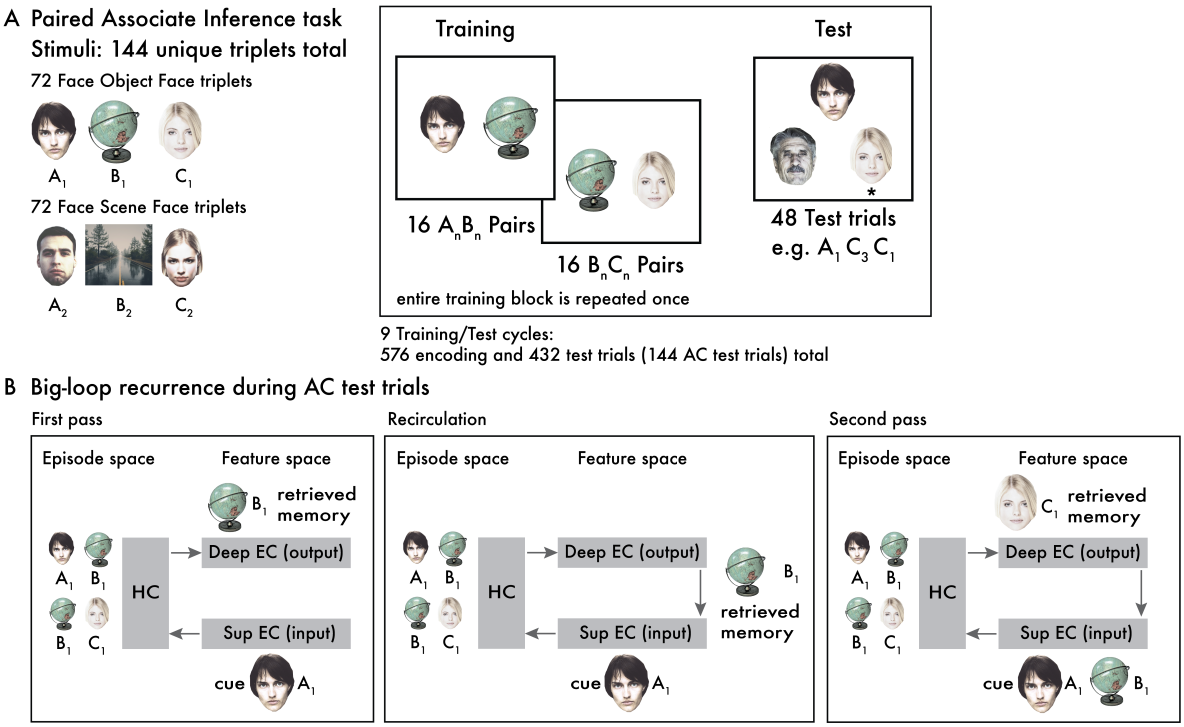

Figure 1. Experimental design and background. A. Experimental design. 144 triplets (i.e. $A_nB_nC_n$, where n=1:144) were used as stimuli in which A and C were always faces, linked by a scene or object (B). After two training exposures (AB and BC pairs presented separately), subjects were tested on direct associations (AB/BC test trials) and indirect associations (AC test trials). During inference test trials (i.e. AC) only faces were displayed, allowing us to investigate reactivation of retrieved memory content (connecting B item) by decoding scenes vs. objects from the neural activation. Note that lures (i.e. C_3 in illustrated test trial) were associated with the same category (object/scene) as the target. The star indicates the correct answer (C_1). Note that no feedback was given. B. Schematic illustration of how big-loop recurrence supports performance on AC test trials. The first pass results in hippocampal memory retrieval and updating of the representation on the output (dEC) layer (i.e. activation of the representation corresponding to B_1). Next, the product of the initial cycle of retrieval is recirculated from the output layer (dEC) to the input layer (sEC). Then in the second pass, the updated input layer representation on the sEC – which consists of the initial query and the product of the previous cycle of retrieval – re-queries the hippocampus. Following subsequent cycles involving an interplay between feature space and episode space, this results in the network producing the correct answer (C_1).

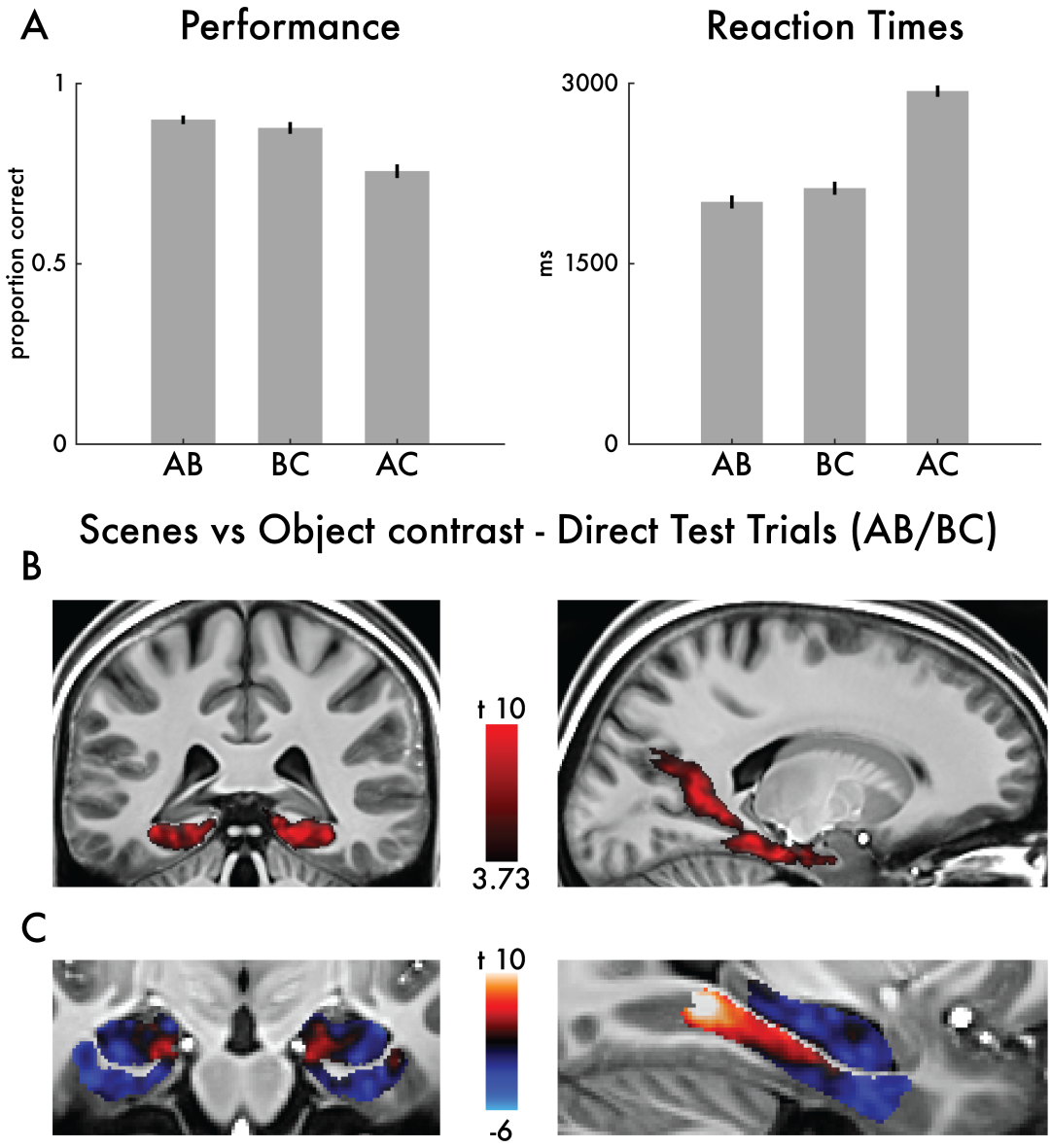

Figure 2. ***Behavioral and univariate results*** A. Behavioral results. Subjects displayed above chance accuracy on AC trials, and show the expected pattern of results, with AC test performance significantly lower than on direct (AB/BC) test trials, and reaction time significantly higher on AC test trials. B. The Scene vs Object group-level contrast in direct Test trials (AB and BC) ($p < 0.001$, uncorrected) spans the posterior MTL to occipital cortex, incorporating both the posterior parahippocampal gyrus and retrosplenial cortex. C. Displaying the Scene vs Object group-level contrast restricted to the medial temporal lobe to visualize a posterior-anterior gradient for scenes and objects respectively; contrast scaled from negative (objects) to positive (scenes), note that the scaling is uneven to display the relatively weaker object activation.

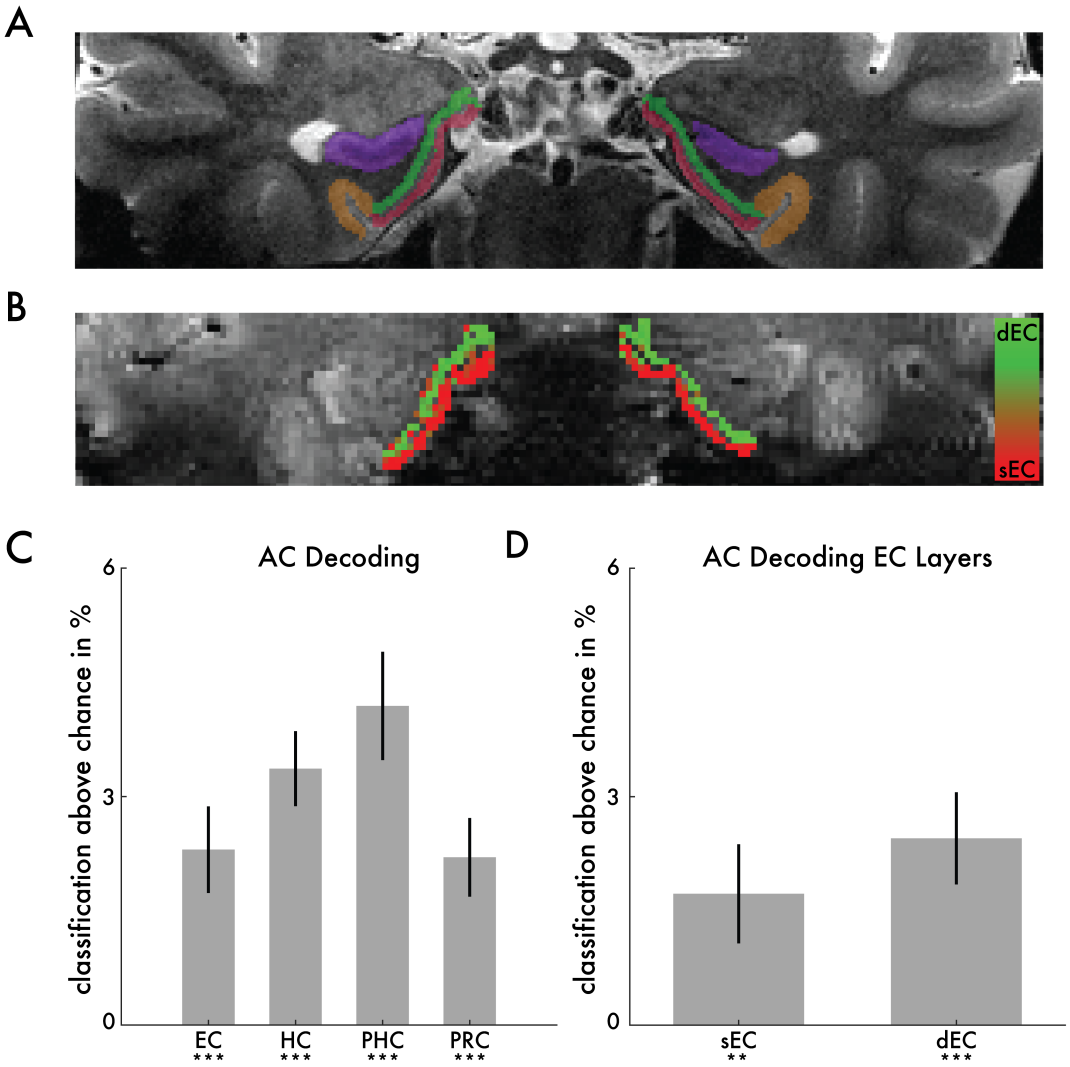

Figure 3. ***AC test decoding results.*** A. Anatomical masks of one example subject displayed on T2 image (red: sEC; green: dEC; purple: Hippocampus; orange: PRC). B. Depiction of classification evidence on a single subject mean EPI, classifying whether a voxel belongs to the sEC or dEC based on its activity profile over trials (red and green indicating maximal evidence for a voxel belonging to the sEC or dEC, respectively; see Fig. S3B for all subjects). The subject with the median classification accuracy across voxels is shown. C. We find reactivation of memory content (scene/object distinction in AC test trials, in which no scene or object was displayed) significantly above chance throughout the MTL (EC: $2.3\% \pm 0.6$, $p < 0.001$; HC: $3.5\% \pm 0.5$, $p < 0.001$; PHC: $4.2\% \pm 0.7$, $p < 0.001$; PRC: $2.2\% \pm 0.5$, $p < 0.001$). D. Crucially, the sEC also shows significant reactivation levels, consistent with information passing back to the entorhinal input layer (sEC: $1.7\% \pm 0.7$, $p = 0.007$; dEC: $2.4\% \pm 0.06$, $p < 0.001$). ** indicates $p < 0.01$, *** indicates $p < 0.001$

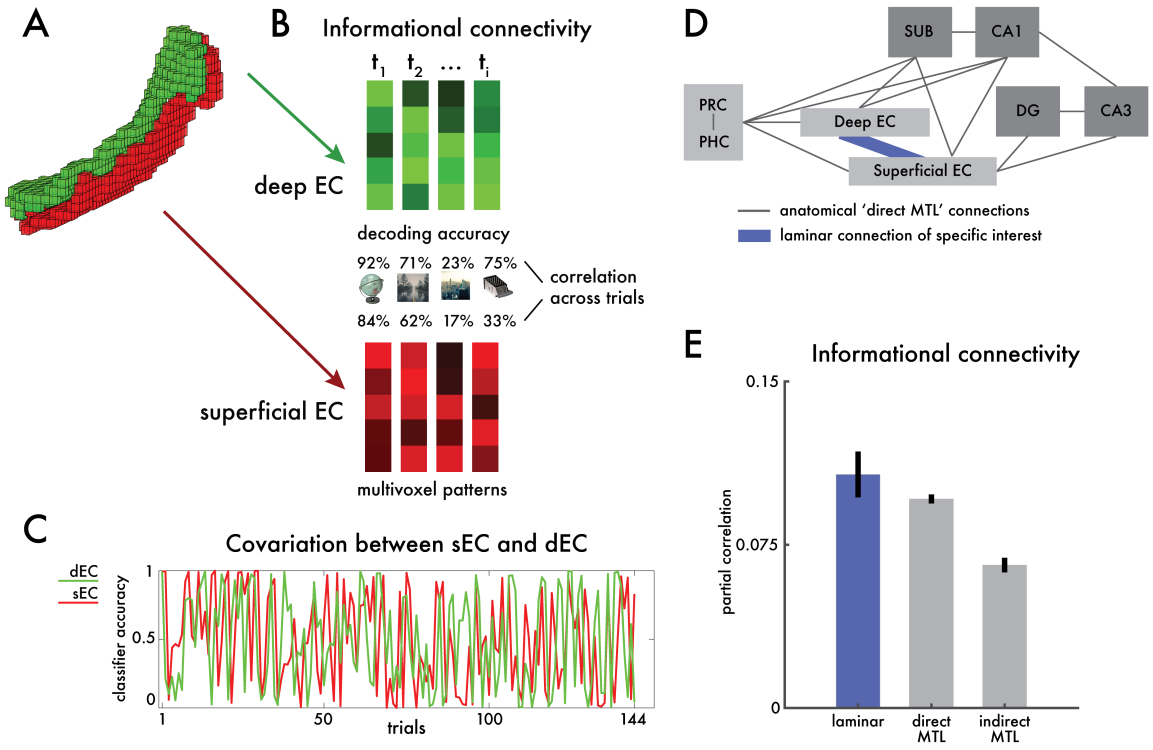

Figure 4. ***Schematic depiction of informational connectivity analysis methods and laminar connectivity results.*** A. Schematic depiction of the sEC and dEC layer separation in EPI space. B. Informational connectivity between two regions is calculated by obtaining the decoding accuracies (scene/object category reactivation) for each trial and region for a given subject and correlating these time series pairwise (adding all other regions' timeseries as covariates). C. Trial by trial covariation between the sEC and dEC information is displayed for one example subject (the subject with median connectivity strength) D. Connectivity-graph of brain regions of interest (Amaral and Lavenex, 2009; Libby et al., 2012). Connectivity patterns were used to divide all possible pairwise connections into 'direct' (connections on the graph) or 'indirect' (off graph) MTL connections, excepting the laminar EC connection which we treat separately. E. Group means of informational connectivity. Error bars display the standard error of the mean. Both the selective laminar connectivity (*mean rho* $= 0.11 \pm 0.01$) and direct MTL (*mean rho* $= 0.096 \pm 0.002$) connectivity are larger ($p < 0.001$, one-tailed) than the indirect MTL connectivity baseline (*mean rho* $= 0.065 \pm 0.003$), suggesting that the laminar connection may reflect a direct flow of information between the EC layers.

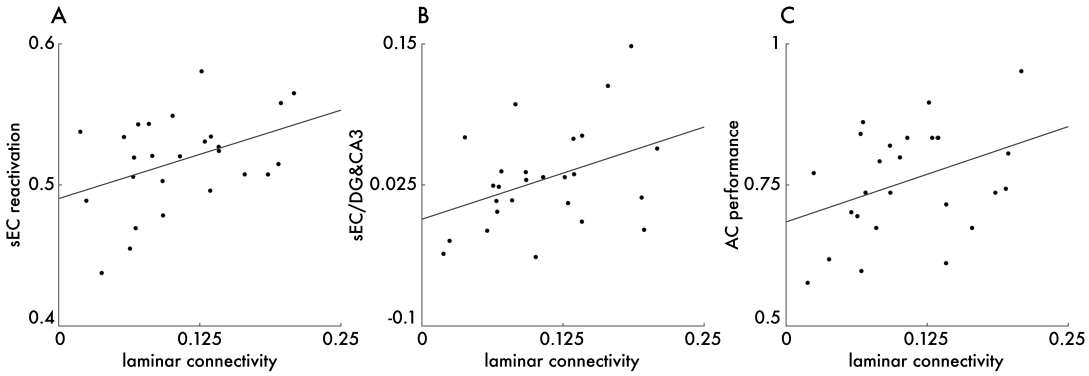

Figure 5. ***Individual variation in entorhinal laminar (sEC/dEC) connectivity correlates with three key variables*** A. Across subjects, laminar connectivity correlates with sEC reactivation of memory content ($r = 0.406$, $p = 0.018$, y axis displays proportion of correctly classified trials, x axis displays averaged connectivity in Spearman's rho in all panels). B. Laminar connectivity correlates with sEC/DG&CA3 connectivity ($r = 0.391$, $p = 0.026$, y axis displays averaged connectivity in Spearman's rho), which reflects the input pathway from the sEC back into the hippocampus. C. Laminar connectivity correlates with Behavioral AC (inference) performance ($r = 0.378$, $p = 0.028$, y axis displays proportion of correct responses). These results support the functional relevance of the entorhinal laminar connection underpinning big-loop recurrence, and its role in reactivation of memory content on the input layer of the EC (sEC), the re-entry of that information into the hippocampus and its relevance for behavioral inference performance. All correlations have one-tailed permuted p-values.

STAR Methods

Contact for Reagent or Resource Sharing

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Raphael Koster (rkoster@google.com).

Experimental Model and Subject Details

Subjects

Twenty-six subjects (recruited via the Otto-von-Guericke University Magdeburg participant database, mean age 27.5$\pm$3.9 years, 12 males) participated in the study and were included in the analysis. Exclusion criteria were metallic implants (other than standard dental implants), tinnitus, known metabolic disorders or a history of neurological or psychiatric disorders. This study was approved by the ethics committee of the Otto-von-Guericke University Magdeburg. All subjects gave written informed consent before participation. A total of four subjects completed the experiment but were excluded from further analysis. Three subjects were excluded due to strong signal dropout in the EC determined by visual inspection of the mean EPI image. One subject was excluded due to poor performance on the paired associate inference task (below chance performance in AC inference trials).

Method Details

Stimuli

The stimuli consisted of 432 digital color photographic images: 72 objects (Bank of Standardized Stimuli, Brodeur et al., 2010), 72 scenes (obtained from various sources on the internet) and 288 faces (Glasgow Unfamiliar Face Database, Burton et al., 2010). All images' mean value was adjusted to the same value. The fixation target was a white fixation cross in the middle of a black screen. Participants watched the stimuli through a mirror mounted on the head coil. Note that for display purposed, depictions of the stimuli displaying faces and scenes in this paper are license-free images obtained from the internet.

Task

During the fMRI session, subjects performed an adapted version of the paired associate inference task (Zeithamova et al., 2012; Banino et al., 2016). For each subject, the stimuli were randomly organized into 144 triplets (ABC; Face-Object-Face (FOF) or Face-Scene-Face (FSF)). Every experimental block utilized 16 different triplets. Each subject completed one scanning session, which contained 3 functional runs. Each functional run contained three experimental blocks (one cycle of training and test utilising 16 triplets – see below). Each block lasted approximately 7 minutes resulting in a total experimental time in the scanner of 63 minutes. Before the scanning, subjects completed a training session consisting of a shortened experimental block. The training was completed once at a self paced speed and once with time-limits exactly like the scanning version of the experiment.

Training block

The training block consisted of being exposed to pairs of stimuli presented next to each other on the screen for 2.5s each followed by a 0.5s fixation screen. First, 16 AB (Face-Object or Face-Scene) pairs were shown in random order. Then, the 16 corresponding BC (Object-Face or Scene-Face) pairs were shown in the order following the initial AB presentation. Then, the entire sequence of 16 AB and 16 BC presentations was repeated in the same order. Which of the two images was presented on the left or right side of the screen was randomized across trials.

Test block

The test block contained three test trials for each of 16 triplets (48 total): direct (i.e. AB and BC) and indirect (AC inference) test trials (see below for details). Each test screen presented one face at the top of the screen (A for AB trials, C for BC trials and A for AC trials) and two images on the bottom of the screen: the target (B for AB trials, B for BC trials and C for AC trials) and the lure. The lure picture was the same scene/object category as the target (or in case of AC trials, associated with the same category) but chosen from a different triplet of the same block. This ensured that the task could not be solved with a category level memory, or with familiarity (i.e. since both choice items were equally familiar). The left-right location of the target and lure was randomized across trials. Subjects were tasked with identifying which of the two bottom pictures was in the same triplet as the picture presented at the top of the screen with a left or right button press.

Solution of the AB and BC test trials can be solved via direct associations, as in each case the pictures were presented together. By contrast, AC trials can only be solved by inferring the connection via the shared linking picture B – hence items A and C were indirectly associated. Importantly, during the AC trials, three faces were displayed on the screen and no object or scene, as A and C were always faces in each triplet. This ensured that any successful classification of objects versus scenes could only be driven by reactivation of the linking image (B). The screen of the test question was displayed for 5s during which subjects could submit their response. After 5s the chosen option was outlined with a green square for 0.5s or, if subjects failed to respond the words 'too slow' were displayed for 0.5s. Each trial was concluded with a fixation screen for 0.5s. The trial order of the 48 test trials was random, but constrained such that for each triplet, the AC test trial had to be presented before the AB or BC trial from the same

triplet. This was done to not contaminate the AC test with learning or exposure from previous trials of the same triplet. The experiment was performed using Presentation software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

Quantification and Statistical Analysis

Behavioral analysis

Behavior was assessed in terms of choice accuracy (i.e. belonging to the same triplet as the stimuli displayed at the top of the screen) and choice reaction time. Differences in these measures were assessed with one way ANOVAs (with the levels AB, BC and AC) and paired t-tests for comparisons of individual conditions. For reaction time, the same procedure was repeated only on correctly answered trials.

F/MRI data acquisition

MRI data were acquired using a 7T MR system (Siemens, Erlangen, Germany). A 32-channel head coil was used (Nova Medical). Before the fMRI session, a whole-head MPRAGE volume (TE=2.8 ms, TR=2500 ms, TI=1050 ms, flip angle=5°, resolution=0.8 mm isotropic, hereinafter 'T1') was acquired. The slices were acquired in an odd-even interleaved fashion oriented parallel to the hippocampus longitudinal axis. EPIs were motion corrected by the sequence built-in algorithm during the online reconstruction (Thesen et al., 2000) and distortion corrected using a point spread function mapping method (In and Speck, 2012). Each subject's fMRI scan consisted of 720 volumes, each comprising 28 T2*-weighted echo planar (hereinafter 'EPI') slices with a resolution of 0.8 x 0.8 mm (TE=22 ms, TR=2000 ms, slice thickness=0.8 mm, FOV=205 mm, matrix=256 x 256, partial Fourier=5/8, parallel imaging with grappa factor 4, bandwidth=1028 Hz/Pixel, echo spacing=1.1 ms, echo train length=40, flip angle=90°) in each of 3 functional runs. After that, the high resolution partial structural volume was acquired (Turbo Spin echo (TSE) with hyperecho (flip angle 60°), TE = 76 ms, TR = 8000 ms, band width = 155 Hz/Pixel, turbo factor 9, 55 slices of 1mm thickness and 10% gap, FOV = 224 mm, matrix = 512 x 512 resulting in an effective spatial resolution of 0.44 mm x 0.44 mm x 1.1 mm, hereinafter 'T2'), with a slice alignment orthogonal to the hippocampal longitudinal axis.

Anatomical ROIs in MTL

Automated segmentation of MTL (ASHS)

For completeness, we briefly describe here the algorithmic scheme of the ASHS software (Yushkevich et al., 2015). Utilizing structural information from both T1 and T2 images, ASHS generates anatomical labels according to a customizable atlas, in which labels were manually defined on the T2 images. For a new, unlabelled brain volume, ASHS evaluates its similarity to each of the labelled brain volumes in the atlas and, using this similarity as weights, produces a probabilistic labelling for the new brain. The final labels of the brain are determined by combining the probabilistic labelling and the gray values in the new brain, using a Bayesian approach. In our case, the atlas consisted of 38 young healthy individuals from another dataset, of which hippocampal subfields (CA1, CA2, CA3, dentate gyrus (DG) and subiculum (SUB)) as well as extrahippocampal subregions (perirhinal (PRC), entorhinal (EC) and parahippocampal cortex (PHC)) were manually segmented and cross-examined by experienced segmentation experts. Importantly, these structural images for atlas building were acquired in the same scanner with exactly the same scanning protocol. In so doing we aimed to maximize the compatibility between the atlas and the current dataset. The labels of the manual segmentation, thus also of the automated segmentation output, include both cortical regions in MTL: PHC, PRC (consisting of A35 and A36) and EC, and hippocampal subfields: CA1, CA2, CA3, DG, SUB. Hippocampal subfield segmentation followed the protocol from Wisse and colleagues (Wisse et al., 2012). Manual delineation of extrahippocampal regions was mainly based on the protocol from Berron and colleagues with a few adaptations, described below (Berron et al., 2017). Delineation of PRC and EC started 3 slices anterior to the hippocampal head. PRC consisted of area 35 and 36 which included the medial and the lateral bank of the collateral sulcus. The posterior border of the PRC and EC was the first slice of the hippocampal body defined by the disappearance of the uncal apex. This was also the starting point for the delineation of the PHC which contained both lateral and medial banks of the collateral sulcus and was segmented until the last slice of the hippocampal tail.

Cortical layer separation

As demonstrated previously (Maass et al., 2014), dividing the EC into 3 layers can capture cytoarchitectural features of the EC. The superficial third (referred to here as 'sEC') of the human EC mainly covers EC layer II and likely parts of EC layer III, projecting into DG/CA3 and CA1, respectively. The deep third (referred to here as 'dEC') mostly covers layer V and VI and is the main output of the HC. The middle layer is excluded from the analysis as functional interpretations are less certain (see Maass et al. (2014) for details).

Our approach to separate the binary mask of EC into 3 different layers is based on shape analysis and cortical thickness estimation. To avoid the aliasing artefacts due to the anisotropic acquisition of the T2 images, we first smoothed the binary mask of EC in T2 space with a Gaussian kernel ($\texttt{FWHM} = 1.1\ mm$), and resampled the smoothed mask to an isotropic resolution of $0.22 \times 0.22 \times 0.22\ mm^3$ (Fig. S6C1-2).

Boundary voxels of the mask in this super-sampled image were identified by using an isosurface method with a volume-preserving threshold, i.e.

$$N(I > threshold) \cdot VS_I = N(BI > 0) \cdot VS_{T2},$$

where I is the smoothed and super-sampled image, BI the binary mask image in T2 space, VS_I and VS_{T2} are their respective voxel volumes.

We further take the smoothed volume as a probabilistic image of the mask, and the gradient vector of this image at each voxel as a potential direction along which the cortical thickness can be estimated. For each boundary voxel, we then estimated the cortical thickness by approximating a line integral along the gradient direction (Fig. S6C3). More specifically, we took sum of dense samples along the direction:

$$thickness(\mathbf{x}) = \int_{\nabla_x} I(\mathbf{x})\,\mathrm{d}x \approx \sum_{i=0}^{N} I(\mathbf{x} + i/N \cdot \nabla_x),$$

where $\mathbf{x}$ is the location of a boundary voxel, and ∇_x the normalized gradient vector at $\mathbf{x}$. In implementation, we used a sampling density of 0.06 *mm*, and took 100 samples along each gradient direction, presuming the maximal thickness < 6 *mm*.

This procedure gave us, for each boundary voxel, an estimate of the cortical thickness in the direction perpendicular to the local boundary surface. Notably, only those voxels on the pial or gray-white matter surfaces provide useful cortical thickness information for laminar separation. We thus removed the voxels residing on the boundaries next to neighbouring regions of grey matter, using an automated approach. Specifically, as shown in Fig. S6C3, voxels falling at these grey matter borders will have much greater estimates of "cortical thickness" than those on the pial or gray-white matter surfaces. We therefore remove all voxels with a cortical thickness estimate greater than $mean(thickness) + 2 \times std(thickness)$ (Fig. S6C4) to automatically detect these outlying values. For the remaining voxels, we started from each boundary voxel, along its gradient direction, to include all the voxels lying within $1/3$ of its thickness to form each segment of the mask. Fig. S6C shows the layer mask after this growing procedure.

Note that despite the illustration in Fig. S6C is in 2D, the actual shape analysis was all done in 3D space, thus avoiding the view-specific limitation of fully manual separation. Furthermore, all anatomical ROIs and masks of separated layers were subsequently examined and, when necessary, corrected manually. Voxels in which two ROIs overlapped after the growing procedure were removed from the mask with lower probability in that voxel. Fig. S6A shows a summary of these ROIs on all subjects.

Cross-modal registration and mapping of ROIs

To precisely extract the layer-specific activity patterns in a cortical area of average thickness ≤ 3 *mm* and measured with imaging resolution of 0.8 *mm*, the separated layer masks of EC in the anatomical image space have to be very accurately transformed into the functional image space. Moreover, due to the partial-coverage scanning protocol, the overlap between the T2 image space, in which the anatomical ROIs were defined, and the functional image space was often too limited to provide reliable registration. Therefore for each subject, the whole brain T1 image was used as a proxy for transforming the ROI masks to the functional image space. Specifically, to obtain the final masks for functional activity extraction, two registration steps were involved: one from the T2 to the T1 and another from the T1 to the functional image.

For every pair of images of each subject, we used the registration tools from 3 software packages: SPM, FSL (Smith et al., 2004) and ANTs (Avants et al., 2008, 2011). Mutual information optimization criteria were used and registration transformation was limited to rigid transform. Field bias of T1 images was corrected using SPM before running registration, to minimize the adverse influence of intensity inhomogeneity on the registration algorithms. In addition to carefully manual examination of the registration results, we set an independent criterion to assess the registration quality objectively: we calculated the agreement between gradient directions at the mask boundaries and that of the target image, after applying the transformation. Here the target image designates the space into which anatomical ROIs were transformed. Based on this boundary gradient agreement criterion, we chose from the 3 registration results the best transformation for mapping the anatomical ROIs to the functional image space.

When mapping the separated layer masks from the super-sampled image space for functional activity pattern extraction, we again adopted a volume-preserving strategy: after smoothing the ROI masks with a Gaussian kernel ($\texttt{FWHM} = 1\ mm$) to a probabilistic mask, we applied the transform matrices and re-thresholded the probabilistic mask to preserve its original binary volume. This strategy also enabled us to resolve cases of partially overlapped voxels by selecting the ROI label with the highest probability. Fig. S6E shows the separated EC layers from the left hemisphere of a typical subject, in the super-sampled space and after transformation into the functional image space.

FMRI data preprocessing

fMRI preprocessing and statistical modeling of the fMRI data was conducted using SPM12 (Wellcome Trust Centre for Neuroimaging, University College London, Penny et al., 2011). Subsequent analyses were run using Matlab. As images were already corrected for distortions and for motion (see "F/MRI data acquisition" section), the preprocessing consisted solely of slice timing correction using default settings. For the multivariate decoding analysis, data were left unsmoothed to preserve the patterns of activation at high spatial resolution. Six volumes that were acquired before the experiment started were discarded. To estimate brain activity for every single trial enabling multi-voxel pattern classification analysis, we computed a separate general linear model (GLM) for each trial (Mumford et al., 2012). For each subject, 1008 GLMs were created, containing the onset of the picture presentation for all training and test trials. Each GLM contains two conditions modelling the scans of one experimental block: the first regressor contained the event (onset of train or test trial) to be modelled in this GLM, the second regressor contained all other events (all other onsets of train or test trials) in the experimental block (one cycle of training and test). Additionally, each model contained 6 regressors of no interest containing motion correction parameters to account for task-related motion. The GLM estimated all voxels contained within a combined mask comprising all anatomical masks of interest created by the automated segmentation (see above). For one subject, the scanner was erroneously stopped before the conclusion of the experiment. For this subject the last 7 trials were excluded. To estimate the activity related to each individual trial, the condition containing the event of interest was contrasted against the condition containing all other events. The resulting t-statistics were used for all multi-voxel pattern classification analyses.

Univariate analysis

For the univariate analysis, a first level model was created modelling 3 sessions, each containing the following conditions (onsets of events with duration 0): 1. Encoding Repetition 1 AB Object 2. Encoding Repetition 1 BC Object 3. Encoding Repetition 1 AB Scene 4. Encoding Repetition 1 BC Scene 5. Encoding Repetition 2 AB Object 6. Encoding Repetition 2 BC Object 7. Encoding Repetition 2 AB Scene 8. Encoding Repetition 2 BC Scene 9. Test AB Object 10. Test BC Object 11. Test AC Object 12. Test AB Scene 13. Test BC Scene 14. Test AC Scene 15. Button Presses 16. Fixation periods. We considered the contrasts; 'AB and BC Test conditions Scene vs Object', 'AC Test conditions Scene vs Object', 'Fixation against the first 14 conditions (picture display)' and 'Test AC trials against baseline (zero over all other conditions)'. The resulting t-contrast images were then normalised to the group template created by using ANTs with all the subjects' T1 whole brain images. The normalised images were then smoothed (Gaussian kernel, $\texttt{FWHM} = 4\ mm$) and group level statistics were computed using the nonparametric toolbox (Nichols and Holmes, 2002). Analyses were restricted to either (a) a mask containing all scanned voxels (the overlap of all subjects' acquisition windows in template space), (b) an anatomical group MTL, or (b) an anatomical group EC mask. The anatomical masks on the group level were created by normalizing each subject's anatomical masks into the ANTS template space, averaging them and binarizing them using a threshold chosen via visual inspection.

To investigate Scene and Object (SvO) representations during visual presentation of scenes and objects we considered a contrast of Test AB Scene and Test BC Scene vs Test AB Object and Test BC Object. The SvO representation of reactivated memory content (linking B item) were investigated with a contrast of Test AC Scene vs Test AC Object. We considered the contrasts masked with scanning coverage of all subject and with the anatomical MTL masks for display purposes. In order to investigate a posterior-anterior Scene to Object gradient within the EC (Litman et al., 2009; Schröder et al., 2015; Berron et al., 2018), we masked each subject's normalized contrast t-image with the group EC mask and collapsed it to one dimension by averaged it to the posterior-anterior axis. We then correlated the average t-value with the voxel position via Pearson correlation, obtaining one correlation value per subject. The inference on the group level was performed via a Wilcoxon sign-rank test. One-tailed significance levels were used due to the directional hypothesis (Berron et al., 2018). We conducted this analysis both for the Test AB/BC SvO contrast as well as the Test AC SvO contrast. We also investigated reinstated memories (Test AC SvO contrast) in a functional region of interest defined by the Test AB/BC SvO contrast. For this, we extracted the average Test AC SvO and Test AC OvS t-value from the top 100 voxels selected for the highest and lowest t-values in the Test AB/BC SvO contrast, respectively. The average t-values were tested against each other with a sign-rank test. Due to the directed hypothesis one-tailed signficance levels were used. To display single subject maps of the SvO and OvS contrasts in AB/BC and AC Test Trials, the contrasts were considered in single subject space and smoothed with a 2mm kernel.

Validating layer separation via classification

In order to assess the presence of layer-specific signals in our data, we took a subject-specific, task-independent approach. Based on the preprocessing steps outlined above, each voxel has a response estimate to every trial in the experiment, which we refer to as the temporal response profile of that voxel. The temporal response profile (i.e. features) used for classification were the 1008 values from the t-images modelling each experimental trial, or the 144 values from the AC test trials. For each subject we took the set of voxels in each of the segmented EC layers (sEC and dEC), and trained a logistic ridge-regression classifier (lambda = 0.01) to classify each voxel as belonging to the

sEC or dEC based on the temporal response profiles. We used a leave-two-out cross-validation approach, where on each fold, one voxel from each layer was held out (note that when necessary random voxels from the larger layer were discarded to equate the sample size in each layer). This classification procedure was repeated for each voxel in both layers. We calculate the percentage of correctly classified (classifier evidence higher than 0.5 for sEC and lower than 0.5 for dEC) voxels for each subject. Group level statistics were calculated with a sign-rank test on accuracy z-scores calculated for each subject, by placing the actual decoding accuracy value across voxels in a distribution of permuted decoding accuracy values. In order to create this distribution of permuted values, we computed the classification of each voxel 100 times with shuffled labels.

Signal intensity, SNR and CNR in EC layers

To assess the overall signal quality in each of the EC layers, we computed and compared a number of metrics. First, we calculated the mean gray value of the mean EPI series of all scans that were acquired during the experiment. Second, tSNR was computed by dividing the mean by the standard deviation of every voxel over the whole timeseries of scans (averaged over the three functional scanning runs separately). Third we considered the univariate contrasts of 'Test AC trials over baseline' and 'picture presentation over fixation', to provide two estimates of contrast-to-noise (CNR, Welvaert and Rosseel, 2013). The average over voxels was then calculated for each region. We directly compare the magnitude of the measures in the sEC and dEC using a sign-rank test.

FMRI MVPA analysis

In order to assess the reactivation of scene and object information, a classifier was trained to distinguish AC trials associated with triplets containing objects and AC trials associated with triplets containing scenes. Note that during AC trials only faces were displayed on the screen, thus any ability to discriminate between the two trial categories must be due to reactivated memory content. For each region of interest (hippocampus, parahippocampal cortex (PHC), perihinal cortex (PRC), deep and superficial layers of the entorhinal cortex (dEC and sEC)) the activation patterns of each AC test trial were extracted. A logistic ridge-regression classifier was used to assess the presence of significant reactivated category content, using a repeated split, 12-fold cross-validation approach. On each data split, the 72 AC Object trials and 72 AC Scene trials were randomly split into 12 folds; each fold contained 6 AC Object and 6 AC Scene trials in the test set and the remaining trials in the training set (for the subject missing trials one fold contained 5 items each). For each fold, a logistic ridge regression (lambda = 0.01) was fitted to the training set, and the fitted model was used to predict an output value for each of the items in the test set (Qian et al., 2013). The output value was passed through a softmax function to produce a probabilistic prediction of the trial category. This prediction was used to calculate 'decoding evidence' which is the likelihood of the trial being classified correctly. This process was repeated 100 times, each time using a new random split of the data. The predictions of the 100 runs were averaged for more stable decoding estimates (Varoquaux et al., 2017). For each subject the decoding evidence was averaged across trials to obtain the overall decoding accuracy. Significance of the decoding accuracy across the population was assessed via a permutation test. The null distribution was obtained for each region for each subject by running the above analyses 1000 times with randomly shuffled trial labels. For each subject a z-score was obtained by scoring the decoding value with respect to the mean and standard deviation of the permuted null distribution. The test assessing

significance on the group level was a sign-rank test over the z-scores of the entire group. For this analysis a one-tailed significance level is used, as one would only expect decoding accuracies above chance. Throughout the main text we report bilateral neural signals by averaging the results of the left and right hemisphere per subject (but see Tables S1 and S2 for results in each hemisphere separately).

Controlling for local vasculature influences

In order to control for the presence of local vasculature influences on the BOLD signal, we ran a control analysis incorporating an additional processing step into the MVPA analysis. We reasoned that, as vasculature artefacts should be spatially local in nature, explicitly removing local correlation between the two layers should minimize their impact. For each voxel in the target layer (e.g. sEC), we found the closest voxel in the opposing layer (e.g. dEC), and regressed the signal out. To ensure that this neighbouring voxel represents the local signal, we smoothed that layer beforehand (smoothing kernel 2mm, constrained to smooth within the structure in order to avoid contamination of the signal through other adjacent structures). As a result, each voxel in the sEC did not contain signal shared with the adjacent dEC voxels, and vice versa. The MVPA analysis described above was then repeated using this processed data.

FMRI informational connectivity analysis

To investigate whether information is passed directly between the deep and superficial layer of the entorhinal cortex, we used an informational connectivity approach (Coutanche and Thompson-Schill, 2013; Aly and Turk-Browne, 2016; Huffman and Stark, 2017). This approach infers informational connectivity from covariation in trial-by-trial decoding accuracy between regions, and is therefore the MVPA analogue of functional connectivity methods commonly applied to univariate data. This technique has been shown to be more sensitive than univariate functional connectivity in a classic dataset of visual processing of objects by Haxby et al. (Haxby et al., 2001; Coutanche and Thompson-Schill, 2013). Selective connectivity between the sEC and dEC (hereafter 'laminar connectivity') was measured as the Spearman correlation between the trial-wise decoding accuracy of the two entorhinal layers, while partialling out the decoding evidence of all other regions of interest in the network (DG, CA1, CA2, CA3, PHC, PRC, SUB, see Zhang et al., 2008; Striem-Amit et al., 2015). Additionally, to rule out that connectivity between layers measured this way is merely an artefact of a general widespread effect in the BOLD response, or spatial correlation between the regions, we added the trial-wise fluctuation in univariate response (the mean t-value of the contrast images over voxels for each trial's estimated activation pattern) in the sEC and dEC as additional covariates. The resulting correlation coefficients were transformed using Fisher's z-transform. The group level inference of whether the correlation coefficients are significantly bigger than zero was based on a sign-rank test (two-tailed). See Fig. 4 for a schematic depiction of the analysis. To validate this approach we also computed the average correlation coefficient strength (pairwise partial Spearman's rho while controlling for all regions of interest and the mean t-value of the two EC layers, as above) of all known direct connections between the MTL regions (all pairwise correlations between the two EC layers, DG, CA1, CA3, PHC, PRC and SUB except the indirect connections) and indirect connections (deep EC layer & DG, deep EC layer & CA3, SUB & DG, SUB & CA3, CA1 & DG, PRC & DG, PRC & CA3, PHC & DG, PHC & CA3, Amaral and Lavenex, 2009; Libby et al., 2012). See Fig. 4D for a depiction of the connectivity graph. If informational connectivity is sensitive to the true underlying anatomical connectivity, we should detect a significant

difference between the direct and indirect connections within the MTL as a whole. If the laminar connection is genuine, it should be significantly stronger than the indirect connections. All of the aforementioned informational connectivity comparisons were conducted on bilateral data, by averaging the informational connectivity across hemispheres for each region (hemisphere specific results are reported in Table S1). All statistical comparisons were conducted using sign-rank tests on the difference between conditions at the group level. Given the directional nature of the hypothesis (connectivity reflecting known anatomical connections) one-tailed significance levels were used when comparing to the indirect baseline. Additionally, we used a classification approach to determine whether the laminar connectivity is classed as 'direct MTL' or 'indirect MTL' connectivity. For this, we trained a classifier (see decoding methods) to distinguish direct and indirect connections for all subjects (one vector with 52 values, the feature being connectivity strength). We then applied the regression equation to the 26 values of the laminar connectivity values. In order to obtain a permuted p-value, we placed the mean value of the resulting classification evidence within a distribution of 10000 values obtained by shuffling the 'direct MTL' and 'indirect MTL labels'.

Local vasculature influences on connectivity

Additional to adding the average BOLD signal in each layers as covariate in the pairwise informational connectivity measures, we performed two control analyses to ensure that the connectivity observed between the layers is not due to a vascular artefact. First, as described above with regard to the MVPA analyses, we removed the local spatial influence of one layer on the other, and recomputed the connectivity on the processed data. In the case of the connectivity analysis, we applied a second method of controlling for local vasculature confounds. Specifically, we subdivided the sEC and dEC masks into lateral and medial portions to avoid correlating adjacent voxels (displayed in Fig. S5). We cut the sEC and dEC into medial and lateral portions by applying a diagonal cut in every coronal slice that aimed to minimize the pairwise size difference between the two medial/lateral sEC and two dEC subregions. We applied the same procedure of decoding Scene/Object reactivation during AC trials in those regions and correlating the decoding accuraries trial-by-trial. This connectivity was always across layer, but now can either be within region (e.g. medial-medial, or lateral-lateral), or across region (medial-lateral). In the latter case, the regions no longer contain any spatially neighbouring voxels across the layers, ensuring that any residual connectivity can not be influenced by local vasculature confounds.

Individual variation analyses

Each subject's measure of "reactivation" was based on the MVPA classification accuracy on the AC inference trials in a given region. Each measure of "connectivity" was based on the informational connectivity between a given pair of regions, as described in an earlier section. Inference performance was each subject's accuracy on the AC test trials, which require inference to solve. Here we were interested in assessing the correlation between certain connectivity and reactivation metrics, as well as with behavioral variation in inference performance, at the level of individual variation. For each such analysis, we used a Pearson correlation (partial correlations when additional variables were controlled for), where the p-value was obtained via its rank in a permuted distribution of 10000 values (Tuğran et al., 2015). Where differences between such correlations are reported, the p-value is based on the rank compared to a permuted distribution of correlation differences. All tests were one sided due to clear directional hypotheses, unless otherwise

stated. All tests were conducted on bilateral data. The key analyses are also reported for the individual hemispheres in Supplemental Materials (Table S2). To establish the expected pattern of results under the big-loop hypothesis, we implemented a simulation of between subject variance in Python, details and results of which are described in S8.