**A map of abstract relational knowledge in human hippocampal-entorhinal cortex**

**Mona M. Garvert1,2, Raymond J. Dolan1,3, Timothy E.J. Behrens1,2**

1Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, UK

2Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford OX3 9D, UK

3Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, Russell Square House, 10-12 Russell Square, London WC1B 5EH, UK

\*Correspondence: [mona.garvert.11@ucl.ac.uk](mailto:mona.garvert.11@ucl.ac.uk)

## Abstract

The hippocampal-entorhinal system encodes a map of space that guides spatial navigation. Goal-directed behaviour outside of spatial navigation similarly requires a representation of abstract forms of relational knowledge. This information relies on the same neural system, but it is not known whether the organisational principles governing continuous maps may extend to the implicit encoding of discrete, non-spatial graphs. Here, we show that the human hippocampal-entorhinal cortex can represent relationships between objects using a metric that depends on associative strength. We reconstruct a map-like knowledge structure directly from a hippocampal-entorhinal fMRI adaptation signal in a situation where relationships are non-spatial rather than spatial, discrete rather than continuous and unavailable to conscious awareness. Notably, the measure that best predicted a behavioural signature of implicit knowledge and BOLD adaptation was a weighted sum of future states akin to the successor representation that has been proposed to account for place and grid-cell firing patterns.

## Introduction

Animals efficiently extract abstract relationships between landmarks, events and other types of conceptual information, often from limited experience. Knowing such regularities can help us act in an environment, because the relationships between items that have never been experienced together can easily be computed and exploited for making novel inference. In physical space, spatially tuned cells in the hippocampal-entorhinal system have precise place (O’Keefe and Dostrovsky, 1971) and grid (Hafting et al., 2005) codes that may form the neural basis of a ‘cognitive map’ (O’Keefe and Nadel, 1978). It is likely that the particular form of these representations enables rapid computations of spatial relationships such as distances and vector paths (Bush et al., 2015; Stemmler et al., 2015). The potential for such rapid online computations embedded into neuronal representations may explain how animals can find novel paths through space (McNaughton et al., 2006; Mittelstaedt and Mittelstaedt, 1980) or rapidly reroute when obstacles are introduced (Alvernhe et al., 2011) or removed (Alvernhe et al., 2008). Indeed, in humans, signals that encode distance metrics between landmarks (Howard et al., 2014; Morgan et al., 2011), and directions to goals (Chadwick et al., 2015) can be read out directly from fMRI data in entorhinal cortex.

The hippocampal formation also encodes non-spatial relationships between objects. When these objects can be laid out in a continuous dimension such as time, hippocampal codes extracted from neuronal ensembles (Rubin et al., 2015) or fMRI voxel patterns (Ezzyat and Davachi, 2014) reflect proximity along this dimension. fMRI signals also appear to reflect veridical angles when two-dimensional abstract spaces are formed from continuous dimensions (Constantinescu et al., 2016; Tavares et al., 2015). However, many relationships that are encoded by the hippocampal formation reflect associations or relationships between discrete objects (Horner et al., 2015; Schapiro et al., 2013, 2012; Wimmer and Shohamy, 2012). To be a useful source of knowledge, many associations must be organised within an associative structure, but it is unclear how such structures might be represented in the absence of a continuous organizing dimension such as space or time.

Highly complex relational structures are often learnt implicitly, i.e. unintentionally and without explicit awareness (Cleeremans et al., 1998; Reber and S., 1989; Seger and Augart, 1994). Neurally, implicitly acquired relational knowledge can be reflected as increases in neural similarity for pairwise associations in the temporal cortex (Schapiro et al., 2012) or for members of a temporal community structure (Schapiro et al., 2013). However, it is unclear whether map or graph-like knowledge structures might be encoded subconsciously, i.e. without subjects being aware of relationships between objects.

Here, we explicitly tested this notion using a functional magnetic resonance imaging (fMRI) adaptation paradigm that allowed us to quantify the relationships between object representations in a neuronal representational space following an implicit learning paradigm. We presented human participants with sequences of objects where stimulus transitions were drawn from random walks along a graph structure. Within entorhinal cortex, a map-like organisation of the relationships between object representations could be extracted from functional magnetic resonance imaging (fMRI) adaptation data acquired on the subsequent day. In this map, associative distance between objects formed a metric that allowed us to extract organising dimensions. This suggests that the brain can automatically organize abstract relational information into map-like structures even if the relationships between objects are non-spatial rather than spatial, discrete rather than continuous and unavailable to conscious awareness. A signature of map-like encoding was also present behaviourally in a separate group of subjects, demonstrating implicit memory of the structure.

We found no evidence for a mapping of discrete relationships into Euclidian space. Instead, the fMRI adaptation pattern as well as behaviour are more consistent with distance measures reflecting the distribution of future states. These principles are consistent with a predictive representation such as the successor representation (Dayan, 1993). Such a predictive map of state space may facilitate the rapid computation of values in a reinforcement learning world (Momennejad et al., 2016; Russek et al., 2016). It has recently been demonstrated that the successor representation can account for a number of properties of place cell and grid cell activity (Stachenfeld et al., 2016, 2014).

## Results

We exposed 23 human participants to object sequences whose stimulus transitions, unbeknownst to them, were determined by a random walk in a graph (Figure 1A). Subjects performed a behavioural cover task, in which they learned to associate a random stimulus orientation with a button press. In the task instructions, any reference to a sequence or an underlying structure was avoided. After the fMRI experiment subjects were debriefed and none reported any explicit knowledge of structure in the task. To test whether this exposure to object sequences induced implicit knowledge about the graph, we scanned the subjects on a subsequent day using fMRI while exposing them to a subset of the same objects presented in a random order (only a reduced graph was presented to increase statistical power, Figure 1B). In 1/10th of the fMRI trials, subjects performed an unrelated cover task, reporting whether a grey patch had been present on the preceding object. Neither accuracy nor response time in this task depended on the object on screen or the transition structure (Figure 1 – Figure supplement 1).

We exploited fMRI adaptation (Barron et al., 2016; Grill-Spector et al., 2006) to investigate the representational similarity for different objects on the graph. We reasoned that in regions encoding a map-like representation of the overall task structure, the degree of similarity in neural representation, and therefore the fMRI adaptation, should decrease as a function of distance between items on the graph. Based on this reasoning we first looked for brain regions whose fMRI response to each object increased as a linear function of the link distance of the preceding item.

This adaptation analysis revealed a peak bilaterally in the entorhinal cortex (Figure 2A, family-wise error-corrected at peak level within a bilateral entorhinal cortex/subiculum mask, left p = 0.014, peak t22= 4.42, [-18, -19, -22] and right p = 0.006, peak t22 = 4.75, [24, -25, -22]. A right, but not the left, peak also survived SVC for a larger region of interest (ROI) comprising the hippocampus, parahippocampal cortex and entorhinal cortex, left p = 0.058 and right p = 0.026, see ROIs in Figure 2 – Figure supplement 1). This adaptation effect cannot be explained by basic statistics of the object sequence, such as the number of times an object occurred during training, or basic features of the graph structure such as the number of neighbours an object has on the graph (Figure 2 – Figure supplement 3).

To confirm the statistical robustness of the effect, and to test whether the effect reflected a gradual increase with distance, we separated the effect into two orthogonal components. These components comprised the difference between connected links (length 1) and all other transitions (lengths 2, 3; Figure 2B, green), and the difference between transitions of length 2 and those of length 3 (Figure 2B, red). These two independent contrasts were used to define ROIs bilaterally in overlapping regions of the entorhinal cortex (both thresholded at p < 0.01 uncorrected; ROI 1: left peak t22= 3.85; [-18, -19, -22] and right peak t22= 3.26; [24, -25, -22], ROI 2: left peak t22= 4.55, [18, -16, -25] and right peak t22= 3.38, [-18, -25 -25]). Because of their statistical independence, we could use the region of interest from one contrast to extract data for the corollary test (t22 = 2.27, p = 0.03 for length 2 vs. length 3 in ROI 1, Figure 2C; and t22 = 2.34, p = 0.03 for connected vs. all other links in ROI 2, Figure 2D). This pair of tests suggest that the fMRI adaptation faithfully represents the link distance. These tests obviate questions of multiple comparisons, because in each case the data is selected from one contrast, and an orthogonal contrast was used for the test statistic.

To further demonstrate this within a single test, we required a coordinate that was independent of all the data. We chose a peak location from an independent study investigating a similar relational measure in spatial maps (Chadwick et al., 2015). Extracting data from this coordinate (ROI 3) revealed a linear effect of link distance (Figure 2E, F2,44 = 10.04, p < 0.001), and correspondingly a significant difference between distances of lengths 1 and 3 (t22 = 3.71, p = 0.001) and lengths 2 and 3 (t22 = 3.19, p = 0.004), but not between distances of lengths 1 and 2 (t22 = 1.67, p = 0.11).

Although this distance effect is suggestive of a map-like organisation, it might also merely reflect the temporal proximity between two objects during training. When the temporal and distance relationships between pairs of objects were allowed to compete for variance in a multiple linear regression, the number of links (t22 = 3.29, p = 0.003), but not time (t22 = 1.27, p = 0.22) explained the neural signal extracted from the independently defined ROI 3 (Figure 3A). Furthermore, relationships between items arranged in a map-like structure are non-directional. Our subjects were not constrained to experience each pair of transitions an equal number of times (Figure 3B). Based upon this we could test whether the fMRI signal was better predicted by the true or symmetrized distance between any two objects. We constructed a measure of the shortest path between each pair of objects according to the actual number of times each transition was experienced by a subject during training (Materials and methods). When allowing this measure to compete with its symmetrized, and thereby non-directional, self in a linear model, it was the symmetrized version alone that predicted the fMRI suppression effect (Figure 3C, t22 = 2.78, p = 0.01 and t22 = -1.64, p = 0.11).

In order to test whether these map-like features are a consequence of a map-like organisation, we organised the signal into a 7x7 matrix with each matrix element reflecting the mean fMRI response across subjects to transitions between the corresponding pairs of objects. For example, element [2,7] in this matrix is the response to object 7 when preceded by object 2 on the graph, averaged across all subjects. Because the signal is suppressed for nearby objects, this matrix is analogous to a distance matrix. When we applied multidimensional scaling to visualise the most faithful 2-dimensional representation of distances in this matrix, the graph structure of our experimental map was recovered despite the subjects’ professed ignorance of any such organisation (Figure 3D). Permutation tests confirm that the MDS-mapped distances are significantly more correlated with link distances of the original graph structure than with link distances of a null distribution consisting of all other complete graphs with 7 links (r = 0.65, p = 0.003, Figure 3 – Figure supplement 2A). Furthermore, no links cross in the graph resulting from the MDS mapping. This is only true for 13.17% of all possible graphs with nodes in the same location, but 7 randomly distributed links (Figure 3 – Figure supplement 2B). Notably, the data were extracted from an independent ROI taken from an experiment investigating maps in allocentric physical space (Chadwick et al., 2015, ROI 3). Results are comparable if parameter estimates are extracted from an anatomically defined region of interest comprising the entorhinal cortex and the subiculum (Figure 3 – Figure supplement 3).

In the reinforcement learning literature it has been suggested that a cognitive map of the relationship between states may be most useful if the representation of a state is predictive in nature and reflects the distribution of likely future states. This idea has been formalized as the successor representation (Dayan, 1993; Momennejad et al., 2016; Russek et al., 2016), proposed to be encoded by hippocampal place cells (Stachenfeld et al., 2016, 2014). According to this view, hippocampal place cells do not encode an animal’s current location in space, but instead encode a predictive representation of future locations. The successor representation may facilitate reinforcement learning, because the resulting predictive measure of future states could be flexibly combined with reward representations to enable rapid computation of navigational trajectories (Baram et al., 2017; Dayan, 1993; Momennejad et al., 2016; Russek et al., 2016).

Mathematically, the successor representation can be computed from the adjacency matrix that defines the relationship between states:

with a discount factor <1. Here, entries for each correspond to the number of possible paths of length between objects and . The successor representation therefore computes the weighted sum of distant future states, with discounted more heavily for larger , i.e. for longer paths between pairs of objects.

Notably this same representation is common in graph theory, where the matrix is termed the matrix resolvent, and is used to measure the proximity or “communicability” between nodes in the graph. Graph theory also proposes a second measure that is closely related, the matrix exponential (Estrada and Hatano, 2008; Estrada and Higham, 2010):

Both measures compute a weighted sum over future states, which easily generalises from continuous to discrete, and from two-dimensional to high-dimensional spaces.

In order to test whether the neural distance effects are consistent with such predictive measures, we tested for areas whose fMRI response to each object increased as a linear function of communicability, corresponding to the negative of the matrix exponential. Compared to the successor representation, the matrix exponential has the advantage that it does not require the fitting of a free parameter. The matrix exponential is small for nodes that are far away from each other on a graph, such that it scales negatively with distance. Unlike a mapping into Euclidian space, communicability significantly distorts the graph structure by shortening links that form part of many paths around the graph structure, and lengthening links that would be less frequently visited by a random navigator (Figure 4A).

We find that neural activity bilaterally in the hippocampal-entorhinal system scales with communicability (Figure 4B,C, family-wise error-corrected at peak level within a bilateral entorhinal cortex/subiculum mask, left p = 0.001, peak t22= 5.47 [-18, -19, -25] and right p = 0.0005, peak t22 = 5.79, [21, -19, -28]). Both clusters also survived SVC for a larger ROI comprising the hippocampus, parahippocampal cortex and entorhinal cortex, left p = 0.004 and right p = 0.004, see ROIs in Figure 2 – Figure supplement 1). Activity in the same areas scales with the negative of the successor representation if the free parameter is set to the commonly used value of ( largest eigenvalue of in modulus, Aprahamian et al., 2016; Benzi and Klymko, 2013, Figure 4 – Figure supplement 1).

In the left hippocampal formation, communicability effects are significant even if Euclidian distances are included as an additional regressor (Figure 4D, p = 0.006, peak t22 = 4.72, [-15, -13, -19], SVC mask 1 and p = 0.027, SVC mask 2). This suggests that the hippocampal-entorhinal system does not map the graph structure into a Euclidian space. Instead, these results are consistent with the view that the distance effect we observe in this system may be a consequence of the hippocampal formation encoding a predictive representation of states within a graph structure (Stachenfeld et al., 2016, 2014).

A map-like representation in the hippocampal-entorhinal cortex suggests that subjects acquired implicit knowledge about the graph structure, even in the absence of explicit awareness of any regularities in the object sequence. To reveal such implicit learning behaviourally we asked an independent group of 26 participants, who were trained in the same way as the scanning cohort on the first graph structure (Figure 5A), to repeat the object orientation cover task on day 2. As was the case for scanned subjects, object transitions were now random and only objects from a reduced graph were presented (Figure 5B). We hypothesized that implicit knowledge about the graph structure would influence response times, such that subjects respond faster if a preceding object in the test sequence is closer on the graph structure underlying the train sequence. Indeed, we found that log-transformed response times were longer the further away the preceding object was on the graph (Figure 5C,D). In line with our imaging results, response times did not scale with link or Euclidian distance between objects, but instead with communicability (communicability: t25 = 2.77, p = 0.01; link distance: t25 = -0.40, p = 0.69, Euclidian distance: t25 = -0.85, p = 0.40, Figure 5C,D).

## Discussion

The hippocampal-entorhinal system is engaged when an animal navigates within a physical environment and acquires flexible knowledge about spatial relationships. In mammals, the hippocampal-entorhinal system contributes to spatial navigation by mapping relationships in situations where knowledge is physical, continuous and consciously available (Chadwick et al., 2015; Derdikman and Moser, 2010; Howard et al., 2014; Spiers and Maguire, 2007). Here, we used a statistical learning paradigm to demonstrate the hippocampal-entorhinal system also efficiently extracts statistical regularities in a non-spatial task where the relationships between items are discrete, and organizes this non-spatial relational knowledge in an abstract relational map, suggesting the hippocampal-entorhinal system creates metric representations of discrete relationships based on associative strengths (Eichenbaum and Cohen, 2014). Notably, there were no continuous dimensions in our discrete stimulus sets. The dimensions had to be extracted from the associations.

These results add to the notion that the hippocampal formation maps experiences across a wide range of different dimensions, thereby supporting flexible behaviour across many domains of life (O’Keefe and Nadel, 1978; Tolman, 1948). Recent studies have focussed on the human hippocampal formation for storing the relational knowledge that make up simple world models in reinforcement learning (Barron et al., 2013; Boorman et al., 2016; Bornstein and Daw, 2013; Wimmer and Shohamy, 2012). Our findings extend these simple associative paradigms to more complex associative maps, and demonstrate that these maps may be learnt implicitly.

Notably, the resulting map is not Euclidian in nature. Instead, we find that the neural data as well as the behaviour are consistent with a measure corresponding to a weighted sum of future states in the graph structure. This is consistent with the idea that reinforcement learning (RL) benefits from knowledge about warped geometries in space, which may be represented as predictive maps, or successor representations, in the hippocampal formation (Stachenfeld et al., 2016, 2014). Such predictive representations of the relationships between discrete objects or states of the world may be combined with reward representations to enable flexible goal-directed behaviour (Baram et al., 2017; Dayan, 1993; Momennejad et al., 2016; Russek et al., 2016).

Our data suggest that the map of the relational structure in discrete abstract space may be encoded by the pattern of firing in the entorhinal cortex, which also encodes maps in continuous physical space. The entorhinal cortex is noted for the presence of grid cells (Hafting et al., 2005), which may provide a metric for measuring distances in physical space (Bush et al., 2015; Stemmler et al., 2015) allowing vector navigation. However, recent theoretical treatments suggest that grid-like firing might also be understood as a principal component decomposition of the covariance information between place cells (Dordek et al., 2016) or the transitions between discrete states (Stachenfeld et al., 2016, 2014). Together with the finding of hippocampal cells in humans that encode individual concepts (that may be analogous to place cells), these theories can explain how grid-like firing patterns could extend to discrete spaces (such as the one we have used here), and also make predictions for how grid-like coding might extend to higher dimensional spaces (Baram et al., 2017).

Our behavioural results are consistent with the observation that the human brain can acquire abstract knowledge unconsciously and automatically by extracting statistical regularities from incidental exposure to events generated according to a set of rules (Berry and Broadbent, 1984; Cleeremans et al., 1998; Reber and S., 1989; Reber, 1967; Seger and Augart, 1994). The resulting implicit knowledge can be used to guide behaviour despite an inability to verbalize the underlying regularities. Neurally, implicit learning of sequences involves medial temporal lobe structures including the hippocampus, subiculum and entorhinal cortex (Schendan et al., 2003). The hippocampal-entorhinal system also responds to the violation of learnt sequence structures (Kumaran and Maguire, 2006) and signals the likelihood of events in learnt sequences (Strange et al., 2005). This may be facilitated by an increase in representational similarity as stimuli become embedded into knowledge structures (Schapiro et al., 2013, 2012). Our results add to this literature by proposing a specific way in which implicit knowledge may be organized in maps to facilitate goal-directed behaviour.

It is notable that we did not find clear evidence for the map-like structure outside the hippocampal formation. It is possible that this is because our study relies on implicit learning, and because the subjects do not have to use the associative structure for any task. Indeed, we note that neural signals can be recorded in frontal and parietal cortex that reflect the ‘state-prediction errors’ that ensue when predicted state relationships are breached during behavioural control (Gläscher et al., 2010). Similar prediction errors in the orbitofrontal cortex during active learning predict later changes in hippocampal representations of the stored model (Boorman et al., 2016). These and similar ideas have led to a theoretical account of place and grid activity in the hippocampal formation as state representations in reinforcement learning models (Stachenfeld et al., 2016, 2014).

It has long been known that the hippocampal formation is important for tasks that rely on associative and relational knowledge. It supports the organization of stimuli across arbitrary stimulus dimensions such as temporal co-occurrence (Schapiro et al., 2013, 2012) or social rank (Kumaran et al., 2012) and organizes behaviourally relevant stimulus categories in a hierarchy (McKenzie et al., 2014). These organizational principles facilitate generalizing over individual episodes (Komorowski et al., 2013) and enable transitive inference by combining newly formed associations between discrete stimuli (Collin et al., 2015; Heckers et al., 2004; Horner et al., 2015; Preston et al., 2004; Schlichting et al., 2015). Value spreading across associated stimulus representations in the hippocampus can then directly influence behaviour in novel decision making situations (Wimmer and Shohamy, 2012). We hope that the current findings help to reconcile these results with the spatial functions of the same neural structures. Such an organization of relational information might be the basis for an animal’s ability to navigate through an abstract concept space and perform flexible computations without direct experience.

## Methods

### Subjects

23 volunteers (aged 18-31, mean age ± std 23.5 ± 3.7 years, 15 males) with normal or corrected-to-normal vision and no history of neurological or psychiatric disorder participated in this experiment. All subjects gave written informed consent and the study was approved by the University College London Hospitals Ethics Committee. The study took place at the Wellcome Trust Centre for Neuroimaging. Subjects were naïve to the purpose of the experiment.

### Stimuli and task

31 coloured and shaded object images which were similar in terms of their familiarity and complexity were selected from the “Snodgrass and Vanderwart ‘Like’ Objects’ picture set (http://wiki.cnbc.cmu.edu/Objects, Rossion and Pourtois, 2004). For each subject, a subset of 12 objects was chosen and randomly assigned to the 12 nodes of the graph shown in Figure 1A. On day 1, subjects were exposed to objects sequences generated from a random walk on the graph, where only objects that were directly connected to an object by a link could follow a presentation of this object. To avoid local repetitions we constrained sequences such that at least 3 objects had to occur between any two presentations of the same object. Each object was randomly presented in one of two orientations, which were mirror images of each other.

Before the start of the experiment, subjects were shown the entire set of 12 stimuli and instructed to remember which of two buttons to press for a particular object orientation (normal or mirrored). During the actual training, subjects were instructed to press the button associated with the stimulus orientation as quickly and accurately as possible while watching the object sequences. Visual feedback after each button press indicated whether a response was correct. Object orientation was randomized across trials, and key assignment was counterbalanced across subjects. Subjects learnt to perform the task quickly and accurately (Figure 1 – Figure supplement 1). Stimuli were presented for 2s and each experimental block consisted of 133 object presentations. Subjects performed this experiment for 12 blocks in total. Between blocks (ca. every 5 min), they were free to take self-paced breaks.

On the next day, subjects were presented with object sequences in the scanner. Only the 7 objects corresponding to the locations illustrated in Figure 1B were presented and stimuli were never repeated. This reduced the total number of stimulus-stimulus transitions and thereby increased statistical power for our key question of interest, as the larger number of times each transition could be probed provided us with a more accurate estimate of the respective suppression effect. Furthermore, stimulus transitions did not follow the graph structure, but were instead randomized with a constraint that each of the 42 possible object transitions occurred exactly 10 times per block (objects were never repeated).

The fMRI experiment consisted of 421 items per run and 3 experimental runs. Stimuli were presented for 1s, with a jittered inter-trial interval (ITI) generated from a truncated Poisson distribution with a mean of 2s. While observing the object sequences subjects performed a cover task of infrequently reporting by button press whether a small grey patch had occurred on a preceding trial. The patch was present on a randomly selected 50% of the objects. Trials on which subjects had to report the existence of a grey patch were signalled by a green cross during the inter-stimulus-interval instead of the standard white cross. The cross was green exactly once after each of the 42 possible transitions, i.e. in 10% of the total number of trials. In 50% of those cases, a patch had been present on the preceding trial. Each correct button press was rewarded with £0.10 paid out in addition to a £33 show-up fee to ensure that subjects attended to the stimuli. Subjects received a brief training on this task before they performed it in the scanner. Key assignment was counterbalanced across subjects. Subjects performed the cover task very well (correct performance rate across subjects: 94% ± 3%, mean ± s.e.m.) confirming that they paid attention to the presented objects throughout the duration of the scan.

After the experiment, subjects were asked whether they noticed any difference between the object sequences presented on day 1 and the object sequences presented in the scanner on day 2. While subjects realized that some objects were missing on day 2, no one reported any awareness of the fact that the sequence differed in any other way. When asked explicitly, no subject was aware of the fact that the sequences on day 1 were generated according to an underlying structure.

### fMRI data acquisition and pre-processing

Visual stimuli were projected onto a screen via a computer monitor. Subjects indicated their choice using an MRI-compatible button box.

MRI data was acquired using a 32-channel head coil on a 3 Tesla Allegra scanner (Siemens, Erlangen, Germany). A T2\*-weighted echo-planar (EPI) sequence was used to collect 43 transverse slices (ascending order) of 2 mm thickness with 1-mm gap and in-plane resolution of 3x3 mm, a repetition time of 3.01 s and an echo time of 70 ms. Slices were tilted by 30° relative to the rostro-caudal axis and a local z-shim with a moment of -0.4mT/m was applied to the OFC region (Weiskopf et al. 2006). The first five volumes of each block were discarded to allow for scanner equilibration. After the experimental sessions, a T1-weighted anatomical scan with 1x1x1 mm resolution was acquired. In addition, a whole-brain field map with dual echo-time images (TE1 = 10 ms, TE2 = 14.76 ms, resolution 3x3x3 mm) was obtained to measure and later correct for geometric distortions due to susceptibility-induced field inhomogeneities.

We performed slice time correction, corrected for signal-bias and realigned functional scans to the first volume in the sequence using a six-parameter rigid body transformation to correct for motion. Images were then spatially normalized by warping subject-specific images to an MNI reference brain, and smoothed using an 8 mm full-width at half maximum Gaussian kernel. All pre-processing steps were performed with SPM12 (Wellcome Trust Centre for Neuroimaging, <http://www.fil.ion.ucl.ac.uk/spm>).

### fMRI data analysis

We implemented three types of event-related general linear models (GLM) to analyse the fMRI data. In the first set of GLMs, we modelled the onset of each of the 7 objects separately to rule out potential confounds associated with subjects’ familiarity with an object or other object-specific activity. This is particularly pertinent because the number of times each object was presented during training depended on its position on the graph due to the random walk along the graph structure (Figure 2 –Figure supplement 3A).

The first set of GLMs contained separate onset regressors for each of the 7 objects with a patch, and without a patch. Each onset regressor was accompanied by different parametric regressors. These corresponded to the link distance (defined as the minimumnumber of linksbetween the pair of items, i.e. distance 1, 2 or 3, Figure 2A), the communicability (see below, Figure 4B) and the negative of the successor representation (Figure 4 – Figure supplement 1) between the object on trial t and the preceding object on trial t-1 on the graph presented in Figure 1B. For the analysis reported in Figure 4D, both communicability and Euclidian distances were included as parametric regressors.

Communicability was computed as the negative of the matrix exponential of the adjacency matrix , describing the relationship between nodes on the graph:

The successor representation was computed as:

with set to the commonly used value of ( largest eigenvalue of in modulus, Aprahamian et al., 2016; Benzi and Klymko, 2013).

Euclidian distances were computed from the graph in Figure 1A, with all distances between objects connected by a link set to 1.

In the second type of GLM (GLM 2), all 42 possible object transitions (object 1 preceded by object 2; object 1 preceded by object 3, …, object 7 preceded by object 6) were modelled separately for patch trials and no-patch trials.

A third type of GLM contained one onset regressor for all objects with a patch, and a separate onset regressor for objects without a patch. Each onset regressor was accompanied by a parametric regressor indicating the number of times an object was presented during training.

All GLMs included a button press regressor as a regressor of no interest. Trials associated with a button press, and the two subsequent trials were not included in the main regressors to avoid button-press related artefacts. All regressors were convolved with a canonical hemodynamic response function. Because of the sensitivity of the BOLD signal to motion and physiological noise, all GLMS also included six motion regressors obtained during realignment as well as ten regressors for cardiac phase, six for respiratory phase and one for respiratory volume extracted with an in-house developed Matlab toolbox as nuisance regressors (Hutton et al., 2011). Models for cardiac and respiratory phase and their aliased harmonics were based on RETROICOR (Glover et al., 2000). Sessions were modelled separately within the GLMs.

The contrast images of all subjects from the first level were analysed as a second-level random effects analysis. We report our results in the hippocampal-entorhinal formation, as this was our *a priori* region of interest, at a cluster-defining statistical threshold of p < 0.001 uncorrected, combined with small volume correction (SVC) for multiple comparisons (peak-level family-wise error (FWE) corrected at p < 0.05). For the SVC procedure we used two different anatomical masks. The first mask consisted of the entorhinal cortex and subiculum alone and was received with thanks from Martin Chadwick (Chadwick et al., 2015, Figure 2 – Figure supplement 1A). The second mask also contained other medial temporal lobe regions implicated in encoding physical space and comprised hippocampus, entorhinal cortex and parahippocampal cortex as defined according to the maximum probability tissue labels provided by Neuromorphometrics, Inc (Figure 2 – Figure supplement 1B). Activations in other brain regions were only considered significant at a level of p < 0.001 uncorrected if they survived whole brain FWE correction at the cluster level (p < 0.05). While no areas survived this stringent correction for multiple comparisons, other regions are reported in Figure 2 – Figure supplement 2 at an uncorrected threshold of p < 0.01 for completeness. While we used masks to correct for multiple comparisons in our region of interest, all statistical parametric maps presented in the manuscript are unmasked.

To independently test a distance-dependent scaling of activity within entorhinal cortex we defined two different regions of interest (ROIs) based on two orthogonal contrasts from GLM 2. The first ROI was defined on the basis of decreased activity in transitions where the preceding object was directly connected with the current object (e.g. regressors corresponding to transition 1-2, 6-4 or 5-7, see Figure 1 – Figure supplement 1C) relative to all other transitions (e.g. regressors corresponding to transition 4-2, 7,4 or 1-7, i.e. [non-connected - connected]). This contrast revealed that clusters in bilateral entorhinal cortex show more adaptation if a preceding object is connected with a currently presented object, relative to a situation where the preceding objects is two or three links away (green, Figure 2B, and Figure 2 – Figure supplement 2B). This defined ROI 1 (thresholded at p < 0.01) from which we then extracted parameter estimates for each of the 42 no-patch transitions and tested for an orthogonal distance effect, namely whether activity differed for transitions with distance 2 relative to transitions of distance 3 using a two-tailed paired t-test.

In a second, independent, test we defined a bilateral entorhinal ROI based on the following contrast: [transitions with 3 links between the relevant objects] - [transitions with 2 links between the relevant objects] for non-patch trials. This contrast is orthogonal to the first contrast, and identified brain regions that responded more strongly on a trial if the preceding object was 3 links, rather than 2 links away (red, Figure 2B and Figure 2 – Figure supplement 2C). Again, we extracted parameter estimates for each of the 42 non-patch onset regressors from this ROI 2, and performed an orthogonal test for distant-dependent scaling by investigating whether activity in this region was also significantly different for directly connected vs. non-connected objects using a two-tailed paired t-test (e.g. transition 1-2, 6-4 or 5-7 versus transition 4-2, 7-4 or 1-7, Figure 1 – Figure supplement 1C).

Note the distance-dependent scaling effects cannot be explained by object-specific differences in activity within these ROIs. While the mean activity for different objects differs slightly, but non-significantly (Figure 3 – Figure supplement 1A,C,E), removing these main effects by subtracting the mean activity for each object before performing the above described analyses does not alter the results (Figure 3 – Figure supplement 1B,D,F).

In a further independent test of a distance-dependent scaling of activity in the hippocampal-entorhinal cortex, we extracted parameter estimates from a region of interest defined based on an independent study investigating the representation of a geocentric goal direction in the entorhinal/subicular region (ROI 3, Chadwick et al., 2015). Specifically, we extracted parameter estimates for the 42 non-patch transitions from the peak voxel reported in his study (MNI coordinates: [−20, −25, −24 ]). This definition of a region of interest was non-biased, and allowed us to test directly test for a distance-dependent scaling of activity. We first performed a repeated-measures ANOVA and post-hoc planned t-tests on the parameter estimates sorted according to distance (Figure 2E). To investigate whether information is organized with respect to the distance relationship or with respect to the average time that passed between the occurrence of two objects during training, we performed a multiple linear regression. In this regression analysis, we included one regressor denoting the distance between object pairs on the graph (1, 2 and 3) and a second regressor accounting for the number of objects that had occurred between any pair of objects and during training. Since the duration of object presentations and the ITI during training were constant, this measure was directly proportional to the time elapsed between the occurrence of the two objects. The dependent variable in the regression analysis was the neural activity for the 42 non-patch transition regressors extracted from this independently defined peak voxel (ROI 3, Figure 3A). To assess the significance across subjects we performed two-tailed paired t-tests on the regression coefficients.

To test for the directionality of the distance effect in the entorhinal cortex, we exploited the fact that subjects were not exposed to transitions between connected objects in the two directions (e.g. 5 followed by 3 vs. 3 followed by 5) equally often. To assess the variability in the number of times a transition was experienced in one vs. the other direction during training, we defined an asymmetry index as

where corresponds to the number of times object y was preceded by object x during training and corresponds to the number of times object x was preceded by object y during training. An asymmetry index of 0 corresponds to a perfect symmetry, i.e. the transition was experienced equally often in both directions, whereas an asymmetry index of 1 corresponds to maximal asymmetry, i.e. the transition was only ever experienced in one direction. Across subjects and transitions there was a large variability in the asymmetry index (Figure 3B).

We could exploit this variability to test for non-directionality in the neural signature, a feature of a map-like structure. We converted the number of times each transition was experienced into a distance measure for each individual subject according to the following equation:

Here, denotes the length of the shortest path between two connected objects. It is computed based on the number of times this particular transition was experienced during training () relative to the number of times the most visited transition was experienced (). The length of the path between objects that were two or three links away was then computed as the single-source shortest path between these objects (by adding the pathlengths for connected objects linking these two objects and choosing the shortest one). To compute the ‘symmetric shortest path measure’, the directional path-length measures (e.g. 5-2 and 2-5) were averaged. The directional, and the symmetric shortest path measures were used as regressors to predict the neural signal extracted from the peak voxel in ROI 3 (Figure 3C). To assess the significance across subjects we performed two-tailed paired t-tests on the regression coefficients.

To visualise the representation of the graph structure in the entorhinal cortex, we performed multi-dimensional scaling (MDS) on the neural activity extracted from the same peak voxel (ROI 3). MDS arranges objects spatially such that the distances between them in space correspond to their similarities as defined by the distance matrix as well as possible. Here, we estimated the configuration of objects in two dimensions using the corresponding inbuilt matlab function mdscale. Specifically, MDS was performed on a matrix denoting the mean neural activity across subjects for each pair of transitions. For example, element 2-5 in the matrix corresponded to the average activity across subjects on trials where object 5 was preceded by object 2 and element 5-2 corresponded to the average activity across subjects on trials where object 2 was preceded by object 5. Because neural activity scales with distance, this matrix effectively corresponds to a distance or similarity matrix. Note that multi-dimensional scaling can only be performed on symmetric matrices with positive entries. We therefore normalized the matrix by subtracting the minimum value of the matrix and adding 1, and then symmetrized it by averaging the top and the bottom triangles.

We tested the map-like representation for significance by comparing the Euclidian distances resulting from projecting our raw data into a 2-dimensional space using multi-dimensional scaling to a null distribution of graph structures generated by permuting the links. Specifically, the null-distribution was generated by keeping the nodes in the location identified by the MDS, but permuting the 7 links making up the graph structure to random nodes. Only complete graphs were included in the null distribution, i.e. graphs where each node was connected to each other node, either directly or indirectly. This results in 68295 unique graphs. We then computed link distances for each graph and correlated the resulting link distance measure with the distances resulting from performing multi-dimensional scaling on the average fMRI response. This provided us with a null distribution we could compare the correlation between the actual graph’s link distance and the MDS-mapped data to.

As a second test of the mapping we compute the number of line crossings in this null distribution. A 2-dimensional map is characterised by the fact that there are no line crossings between pairs of directly connected nodes. In the null-distribution, this is only true for 13.17% of all graphs.

We repeated all analyses reported in Figure 3 for parameter estimates extracted from an anatomically-defined region of interest comprising the entorhinal cortex and subiculum (Figure 3 – Figure supplement 3).

**Behavioural experiment**

A separate group of 26 subjects participated in a behavioural version of the experiment. Day 1 of the behavioural experiment was designed to be exactly equivalent to day 1 of the fMRI experiment, with subjects performing 10 (n = 14) or 12 (n = 12) blocks of the object orientation cover task on object sequences generated according to a random walk along the graph structure. On day 2, subjects performed the same object orientation cover task on the reduced set of objects presented to subjects participating in the fMRI experiment in the scanner. Trial transitions were pseudo-randomised to ensure that each object was preceded by each other object the same number of times. This enabled us to test for changes in response times (RTs) with distance between objects on the graph. Subjects performed 10 blocks of test trials with self-paced breaks in between blocks, with 127 objects presented in each block. Thereby, each stimulus-stimulus transition was probed 3 times per block, or 30 times across the experiment as a whole.

All analyses were performed on log-transformed response times to normalise response time measures. To account for object-specific effects that are independent from any distance-dependent scaling, we subtracted the mean response time for each object and subsequently computed average demeaned response times for each of the 42 stimulus-stimulus transitions per block (object 1 preceded by object 2; object 1 preceded by object 3, …, object 7 preceded by object 6). We averaged these measures across blocks to obtain one representative measure per transition and subject.

To test for scaling of response times with distance on the graph, we regressed communicability, link distance and Euclidian distance for each of the 42 transitions onto subject-specific response time measures. The significance of the regression across subjects was assessed using two-tailed paired t-tests on the resulting regression coefficients (Figure 5C). To visualise the relationship between communicability and response times, we sorted the data according by communicability, created 7 bins with an equal number of transitions in each bin, and plotted the mean log response times across subjects for each bin (Figure 5D).

**Acknowledgements**

## We thank Alon Baram, Tim Muller, Zeb Kurth-Nelson, Philipp Schwartenbeck, Alexandra Constantinescu, Archy de Berker and Helen Barron for discussions and comments on the manuscript. This work was supported by the Wellcome trust (4-year PhD studentship 097267/Z/11/Z to M.M.G, Senior Investigator Award to R.J.D., 098362/Z/12/Z, and Senior Research Fellowship104765/Z/14/Z to T.E.J.B), the Joint Initiative on Computational Psychiatry and Ageing Research between the Max Planck Society and University College London (R.J.D.) and the James S. McDonnell Foundation (JSMF220020372, T.E.J.B.). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust (Strategic Award Grant 091593/Z/10/Z).

## References

Alvernhe, A., Save, E., Poucet, B., 2011. Local remapping of place cell firing in the Tolman detour task. Eur. J. Neurosci. 33, 1696–705. doi:10.1111/j.1460-9568.2011.07653.x

Alvernhe, A., Van Cauter, T., Save, E., Poucet, B., 2008. Different CA1 and CA3 representations of novel routes in a shortcut situation. J. Neurosci. 28, 7324–33. doi:10.1523/JNEUROSCI.1909-08.2008

Aprahamian, M., Higham, D.J., Higham, N.J., 2016. Matching exponential-based and resolvent-based centrality measures. J. Complex Networks 4, 157–176. doi:10.1093/comnet/cnv016

Baram, A.B., Muller, T.H., Behrens, T.E.J., 2017. Intuitive planning: global navigation of cognitive maps with grid-like global representations. Proc. Comput. Syst. Neurosci. Meet.

Barron, H.C., Dolan, R.J., Behrens, T.E.J., 2013. Online evaluation of novel choices by simultaneous representation of multiple memories. Nat. Neurosci. 16, 1492–1498. doi:10.1038/nn.3515

Barron, H.C., Garvert, M.M., Behrens, T.E.J., 2016. Repetition suppression: a means to index neural representations using BOLD? Philos. Trans. R. Soc. B Biol. Sci. 371, 20150355. doi:http://dx.doi.org/10.1098/rstb.2015.0355

Benzi, M., Klymko, C., 2013. Total communicability as a centrality measure. J. Complex Networks 1, 124–149. doi:10.1093/comnet/cnt007

Berry, D.C., Broadbent, D.E., 1984. On the relationship between task performance and associated verbalizable knowledge. Q. J. Exp. Psychol. Sect. A 36, 209–231. doi:10.1080/14640748408402156

Boorman, E.D., Rajendrana, V.G., O’Reilly, J.X., Behrens, T.E.J., Rajendran, V.G., O’Reilly, J.X., Behrens, T.E.J., Rajendrana, V.G., O’Reilly, J.X., Behrens, T.E.J., 2016. Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. Neuron 89, 1343–1354. doi:10.1016/j.neuron.2016.02.014

Bornstein, A.M., Daw, N.D., 2013. Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. PLoS Comput. Biol. 9, e1003387. doi:10.1371/journal.pcbi.1003387

Bush, D., Barry, C., Manson, D., Burgess, N., 2015. Using grid cells for navigation. Neuron 87, 507–520. doi:10.1016/j.neuron.2015.07.006

Chadwick, M.J., Jolly, A.E.J., Amos, D.P., Hassabis, D., Spiers, H.J., 2015. A goal direction signal in the human entorhinal/subicular region. Curr. Biol. 25, 87–92. doi:10.1016/j.cub.2014.11.001

Cleeremans, A., Destrebecqz, A., Boyer, M., 1998. Implicit learning: news from the front. Trends Cogn. Sci. 2, 406–416. doi:10.1016/S1364-6613(98)01232-7

Collin, S.H.P., Milivojevic, B., Doeller, C.F., 2015. Memory hierarchies map onto the hippocampal long axis in humans. Nat. Neurosci. 18, 1562–1564. doi:10.1038/nn.4138

Constantinescu, A.O., O’Reilly, J.X., Behrens, T.E.J., 2016. Organizing conceptual knowledge in humans with a grid-like code. Science 352, 1464–8. doi:10.1126/science.aaf0941

Dayan, P., 1993. Improving generalization for temporal difference learning: The successor representation. Neural Comput. 5, 613–624. doi:10.1162/neco.1993.5.4.613

Derdikman, D., Moser, E.I., 2010. A manifold of spatial maps in the brain. Trends Cogn. Sci. 14, 561–9. doi:10.1016/j.tics.2010.09.004

Dordek, Y., Soudry, D., Meir, R., Derdikman, D., 2016. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. Elife 5, e10094. doi:10.7554/eLife.10094

Eichenbaum, H., Cohen, N.J., 2014. Can we reconcile the declarative memory and spatial navigation views on hippocampal function? Neuron 83, 764–770. doi:10.1016/j.neuron.2014.07.032

Estrada, E., Hatano, N., 2008. Communicability in complex networks. Phys. Rev. E 77, 036111. doi:10.1103/PhysRevE.77.036111

Estrada, E., Higham, D.J., 2010. Network Properties Revealed through Matrix Functions. SIAM Rev. 52, 696–714. doi:10.1137/090761070

Ezzyat, Y., Davachi, L., 2014. Similarity breeds proximity: pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. Neuron 81, 1179–89. doi:10.1016/j.neuron.2014.01.042

Gläscher, J., Daw, N., Dayan, P., O’Doherty, J.P., 2010. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron 66, 585–595. doi:10.1016/j.neuron.2010.04.016

Glover, G.H., Li, T.Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44, 162–167.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10, 14–23. doi:10.1016/j.tics.2005.11.006

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., Moser, E.I., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806. doi:10.1038/nature03721

Heckers, S., Zalesak, M., Weiss, A.P., Ditman, T., Titone, D., 2004. Hippocampal activation during transitive inference in humans. Hippocampus 14, 153–62. doi:10.1002/hipo.10189

Horner, A.J., Bisby, J.A., Bush, D., Lin, W.-J., Burgess, N., 2015. Evidence for holistic episodic recollection via hippocampal pattern completion. Nat. Commun. 6, 7462. doi:10.1038/ncomms8462

Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M., Staskute, L., Spiers, H.J., 2014. The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. Curr. Biol. 24, 1331–1340. doi:10.1016/j.cub.2014.05.001

Hutton, C., Josephs, O., Stadler, J., Featherstone, E., Reid, A., Speck, O., Bernarding, J., Weiskopf, N., 2011. The impact of physiological noise correction on fMRI at 7 T. Neuroimage 57, 101–112. doi:10.1016/j.neuroimage.2011.04.018

Komorowski, R.W., Garcia, C.G., Wilson, A., Hattori, S., Howard, M.W., Eichenbaum, H., 2013. Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts. J. Neurosci. 33, 8079–87. doi:10.1523/JNEUROSCI.5458-12.2013

Kumaran, D., Maguire, E.A., 2006. An unexpected sequence of events: Mismatch detection in the human hippocampus. PLoS Biol. 4, 2372–2382. doi:10.1371/journal.pbio.0040424

Kumaran, D., Melo, H.L., Duzel, E., 2012. The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. Neuron 76, 653–666. doi:10.1016/j.neuron.2012.09.035

McKenzie, S., Frank, A., Kinsky, N., Porter, B., 2014. Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. Neuron 83, 202–215.

McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.-B., 2006. Path integration and the neural basis of the “cognitive map”. Nat. Rev. Neurosci. 7, 663–78. doi:10.1038/nrn1932

Mittelstaedt, M.-L., Mittelstaedt, H., 1980. Homing by path integration in a mammal. Naturwissenschaften 67, 566–567. doi:10.1007/BF00450672

Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N., Gershman, S.J., 2016. The successor representation in human reinforcement learning. bioRxiv.

Morgan, L.K., Macevoy, S.P., Aguirre, G.K., Epstein, R.A., 2011. Distances between real-world locations are represented in the human hippocampus. J. Neurosci. 31, 1238–45. doi:10.1523/JNEUROSCI.4667-10.2011

O’Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 34, 171–175. doi:10.1016/0006-8993(71)90358-1

O’Keefe, J., Nadel, L., 1978. The hippocampus as a cognitive map. Oxford: Clarendon Press.

Preston, A.R., Shrager, Y., Dudukovic, N.M., Gabrieli, J.D.E., 2004. Hippocampal contribution to the novel use of relational information in declarative memory. Hippocampus 14, 148–52. doi:10.1002/hipo.20009

Reber, A.S., 1967. Implicit learning of artificial grammars. J. Verbal Learning Verbal Behav. 6, 855–863. doi:10.1016/S0022-5371(67)80149-X

Reber, A.S., S., A., 1989. Implicit learning and tacit knowledge. J. Exp. Psychol. Gen. 118, 219–235. doi:10.1037/0096-3445.118.3.219

Rossion, B., Pourtois, G., 2004. Revisiting Snodgrass and Vanderwart’s object pictorial set: the role of surface detail in basic-level object recognition. Perception 33, 217–236.

Rubin, A., Geva, N., Sheintuch, L., Ziv, Y., 2015. Hippocampal ensemble dynamics timestamp events in long-term memory. Elife 4. doi:10.7554/eLife.12247

Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J., 2016. Predictive representations can link model - based reinforcement learning to model - free mechanisms. bioRxiv. doi:10.1101/083857

Schapiro, A.C., Kustner, L. V., Turk-Browne, N.B., 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr. Biol. 22, 1622–1627. doi:10.1016/j.cub.2012.06.056

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M., 2013. Neural representations of events arise from temporal community structure. Nat. Neurosci. 16, 486–492. doi:10.1038/nn.3331

Schendan, H.E., Searl, M.M., Melrose, R.J., Stern, C.E., 2003. An fMRI Study of the Role of the Medial Temporal Lobe in Implicit and Explicit Sequence Learning. Neuron 37, 1013–1025. doi:10.1016/S0896-6273(03)00123-5

Schlichting, M.L., Mumford, J.A., Preston, A.R., 2015. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. Nat. Commun. 6, 8151. doi:10.1038/ncomms9151

Seger, C.A., Augart, C., 1994. Implicit learning. Psychol. Bull. 115, 163–196. doi:10.1037/0033-2909.115.2.163

Spiers, H.J., Maguire, E.A., 2007. A navigational guidance system in the human brain. Hippocampus 17, 618–26. doi:10.1002/hipo.20298

Stachenfeld, K.L., Botvinick, M., Gershman, S.J., 2014. Design principles of the hippocampal cognitive map, in: Advances in Neural Information Processing Systems. pp. 2528–2536.

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J., 2016. The hippocampus as a predictive map. bioRxiv.

Stemmler, M., Mathis, A., Herz, A.V.M., 2015. Connecting multiple spatial scales to decode the population activity of grid cells. Sci. Adv. 1, e1500816–e1500816. doi:10.1126/science.1500816

Strange, B.A., Duggins, A., Penny, W., Dolan, R.J., Friston, K.J., 2005. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Networks 18, 225–230. doi:10.1016/j.neunet.2004.12.004

Tavares, R.M.M., Mendelsohn, A., Grossman, Y., Williams, C.H.H., Shapiro, M., Trope, Y., Schiller, D., 2015. A map for social navigation in the human brain. Neuron 87, 231–243. doi:10.1016/j.neuron.2015.06.011

Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol. Rev. 55, 189–208. doi:10.1037/h0061626

Wimmer, G.E., Shohamy, D., 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338, 270–3. doi:10.1126/science.1223252

## Figure legends

**Figure 1. Experimental design. A** Graph structure, used to generate stimulus sequences on day 1. Trial transitions were drawn from random walks along the graph. **B** Objects on reduced graph, presented to subjects in the scanner on day 2. Trial transitions were random. In both sessions, participants performed simple behavioural cover tasks. See Figure 1 - Figure supplement 1 for behavioural performance during the training and the scan sessions.

**Figure 2. fMRI adaptation in the hippocampal-entorhinal cortex decreases with distance on the graph. A** Whole-brain analysis showing a decrease in fMRI adaptation with link distance in the hippocampal-entorhinal cortex, thresholded at p < 0.01, uncorrected for visualisation. **B** Within hippocampal-entorhinal cortex, green indicates greater suppression if the preceding stimulus was a neighbour relative to a stimulus two or three links away. Red indicates greater suppression if a preceding stimulus was two links away than three links away. The depicted areas were used as regions of interest for analyses in c (green) and d (red). **C** Parameter estimates for link 2 vs link 3 transitions extracted from the green entorhinal ROI in Figure 2A (t22 = 2.27, p = 0.03). Other brain areas do not show this increase in activity with distance (Figure 2 – Figure supplement 2). **D** Parameter estimates extracted from the red entorhinal ROI in Figure 2A, sorted according to whether objects were connected on the graph or not (t22 = 2.34, p = 0.03). **E** Parameter estimates extracted from the peak MNI coordinate reported in Chadwick et al. (2015), [−20, −25, −24] and sorted according to distance (F2,44 = 10.04, p = 0.0003). See Figure supplement 1 for masks used for small-volume correction in Figure 2A, Figure supplement 2 for distance-dependent scaling effects in other brain regions and Figure supplement 3 for effects of object familiarity and centrality. Error bars show mean and s.e.m.; a.u., arbitrary units.

**Figure 3. Relational information is organized as a map. A** Linear regression on neural activity with number of links and average time between two objects during training as regressors (t22 = 3.29, p = 0.003 and t22 = 1.27, p = 0.22). **B** Absolute difference in the number of times a transition was visited in one vs. the other direction (e.g. 5 preceded by 1 vs. 1 preceded by 5) for all subjects. **C** Multiple linear regression on neural activity with the shortest path between objects, and the symmetrized shortest path between objects as regressors (t22 = -1.64, p = 0.11 and t22 = 2.78, p = 0.01). **D** 7x7 matrix representing the average fMRI signal in response to an object depending on which other object preceded, averaged across subjects and symmetrized. Objects were never repeated during scanning, the diagonal entries are therefore set to 0. This matrix was used for the MDS visualised in **E**. **E** Visualisation of the localization of the object representations in a 2-dimensional space according to multiple-dimensional scaling. Lines indicate transitions experienced during training. The distances between the resulting locations of nodes in a 2-D space are significantly correlated with link distances of the original graph structure (r = 0.65, p = 0.0003, see Figure supplement 2 for the null distribution used for the permutation test). All analyses were performed on dataextracted from the peak MNI coordinate reported in Chadwick et al. (2015), [−20, −25, −24], but also hold in an anatomically defined region of interest including the entorhinal cortex and the subiculum (Figure supplement 3). See Figure supplement 1 for results if object-specific activity is removed. Error bars show mean and s.e.m.; a.u., arbitrary units.

**Figure 4. fMRI adaptation in the hippocampal-entorhinal system is consistent with predictive representations of relational knowledge. A** Visualisation of communicability coordinates for the graph structure by performing multi-dimensional scaling on the communicability matrix. **B** Whole-brain regression of communicability onto neural activity. **C** Visualisation of the communicability effect. Average parameter estimate across subjects extracted from a bilateral region of interest in B (thresholded at p < 0.01). The colour of the dots corresponds to link distances. This graph is added for visualisation purposes only as the parameter selection is biased. **D** Whole-brain regression of communicability onto neural activity when Euclidian distances are included as an additional regressor in the GLM. All statistical maps are thresholded at p < 0.01 for visualisation.

**Figure 5. Response times reflect graph structure. A** Graph structure, used to generate stimulus sequences on day 1. Trial transitions were drawn from random walks along the graph structure. **B** Objects on reduced graph, presented to subjects on day 2. Trial transitions were random. In both sessions, participants performed an object orientation cover task under which response times were measured. **C** A regression of communicability, link distance and Euclidian distance onto log-transformed response times across subjects (communicability: t25 = 2.77, p = 0.01; link distance: t25 = -0.40, p = 0.69, Euclidian distance: t25 = -0.85, p = 0.40). **F** Visualisation of the relationship between communicability and log-transformed response times. Communicability measures were divided into six bins with an equal number of object-object transitions per bin. The y-axis corresponds to the average demeaned log-response time across subjects for each bin.

**Figure 1 – Figure supplement 1. Task performance. A** Response times and **B** performance on the orientation judgment cover task performed during training on day 1 for each of the twelve blocks. **C** Graph structure indicating the object position. **D** Response time (F6,132 = 0.68, p = 0.67) and **E** percent correct (F6,132 = 0.56, p = 0.77) on the patch detection cover task performed during the fMRI experiment does not differ for the different object locations. **F** Response times on the patch detection cover task does not depend on the distance between objects on the graph (F2,44 = 0.72, p = 0.49). Error bars show mean and s.e.m.; a.u., arbitrary units

**Figure 2 – Figure supplement 1. Anatomically defined regions of interest used for small-volume correction. A** Mask comprising bilateral entorhinal cortex and subiculum, received with thanks from Martin Chadwick (Chadwick et al., 2015). **B** Mask comprising bilateral entorhinal cortex, hippocampus and parahippocampal cortex.Regions were defined using the maximum probability tissue labels provided by Neuromorphometrics, Inc (http://Neuromorphometrics.com)

**Figure 2 – Figure supplement 2. Distance-dependent scaling of neural activity is specific to the hippocampal-entorhinal system. A** All areas displaying a decrease in fMRI adaptation with graph distance. A cluster in subgenual cortex did not survive whole-brain correction for multiple comparisons (punc = 0.0008, pFWE = 0.99, peak t22 = 3.58 [3, 23, -10]). **B** All areas displaying suppression for connected stimuli relative to non-connected stimuli on the graph. In addition to the hippocampal-entorhinal clusters reported in the main text, a cluster in orbitofrontal cortex (OFC, peak t22 = 2.93, [30, 41, -10]) and subgenual cortex (peak t22 = 3.33, [0, 23, -7]) were used to define ROIs to test distance-dependent scaling, see **D,E**. **C** All areas displaying greater suppression if a preceding stimulus was two links away than three links away. The corollary test for a difference in activity for connected versus non-connected objects was only significant in the hippocampal-entorhinal ROI, see Figure 2B, D. **D** Parameter estimates for link 2 vs link 3 transitions extracted from the orbitofrontal cortex ROI in **B**. The difference is not significant (t22 = 0.85, p = 0.41). **E** Parameter estimates for link 2 vs link 3 transitions extracted from the subgenual cortex ROI in **B**. The difference is not significant (t22 = 1.29, p = 0.21). **A-C** are thresholded at p < 0.01 for visualisation. Error bars show mean and s.e.m.; a.u., arbitrary units.

**Figure 2 – Figure supplement 3. Effects of object familiarity. A** Number of times each object was presented during training, averaged across participants.Note that this measure is directly related to the number of neighbours an object has on the graph.During scanning/behavioural testing, each object of the reduced graph was presented equally often (180 times). Error bars denote standard deviation. **B** Decrease of activity with an increase in the number of times an object was presented during training (GLM 3), thresholded at p < 0.01 for visualisation.No cluster survives correction for multiple comparisons. No effect can be detected when testing for a positive effect.

**Figure 3 – Figure supplement 1. The distance-dependent scaling cannot be driven by a main effect of object position.** Position-specific activity in the ROI defined according to **A** a connected < non-connected contrast (ROI 1, Figure 2B, green, F6,132 = 0.88, p = 0.5), **B** a link 2 < link 3 contrast (ROI 2, Figure 2B, red, F6,132 = 1.96, p = 0.08) and **E** the peak coordinate in Chadwick et al. (2015), (ROI 3, F6,132 = 1.9, p = 0.09). The tests reported in Figure 2**C-E** are also significant if performed after removing object-specific activity from the neural data. This was achieved by subtracting the mean activity for each object before testing for **B** a difference in activity for link 2 vs. link 3 transitions in ROI 1 (t22 = 2.10, p = 0.048), **D** a difference in activity for connected vs. non-connected stimuli in ROI 2 (t22 = 2.39, p = 0.03) or **F** a distance effect in ROI 3 (F2,44 = 6.68, p = 0.003).Error bars show mean and s.e.m.; a.u., arbitrary units

**Figure 3 – Figure supplement 2. Map characteristics in a null distribution generated by permuting the links making up the graph structure. A** Correlation of link distances for graphs making up the null distribution with distances resulting from performing multi-dimensional scaling on the neural data. The null distribution was constructed by permuting the 7 links making up the reduced graph structure to a random set of nodes. The correlation between link distances for the actual graph structure and distances resulting from multi-dimensional scaling on the neural data is 0.65. Only 0.28% of graphs show a correlation with the mapped data of 0.65 or higher. **B** Number of line crossings for graphs characterised by nodes in the same location as given by the MDS procedure, but with the 7 links randomly distributed between nodes. Only 13.17% of all possible graphs generated in this way are characterised by 0 line crossings. 0.25 % of all graphs have 0 line crossing and a correlation of 0.65 or higher. The arrows indicate the corresponding measure for the actual data.

**Figure 3 – Figure supplement 3. Distance effects in an anatomically defined region of interest comprising the entorhinal cortex and the subiculum.** **A** Activity scales linearly with link distance (F2,44 = 8.41, p = 0.0008). *Post-hoc* pairwise comparisons revealed a significant difference between distances of lengths 1 and 2 (t22 = 2.36, p = 0.03), lengths 1 and 3 (t22 = 3.51, p = 0.002), and lengths 2 and 3 (t22 = 2.23, p = 0.04). **B** In a multiple linear regression, link distance, but not time, predicts the neural activity (t22 = 2.13, p = 0.04 and t22 = 1.23, p = 0.23). **C** In a multiple linear regression with symmetrised and non-symmetrised distance measures competing for variance, symmetrised distance rather than non-symmetrised distance explains the neural signal (directional distance: t22 = -1.18, p = 0.25, symmetrised distance: t22 = 2.45, p = 0.02). **D** The graph structure can be recovered by performing multi-dimensional scaling on the average fMRI data across subjects (r = 0.66, p = 0.008, permutation test). Data for all analyses are extracted from the anatomically defined region of interested shown in Figure 2 – Figure supplement 1A. Error bars show mean and s.e.m.; a.u., arbitrary units

**Figure 4 – Figure supplement 1. Activity in the hippocampal-entorhinal system is consistent with the successor representation. A** Visualisation of successor representation coordinates for the graph structure by performing multi-dimensional scaling on the negative of the successor representation, with the free parameter set to (= largest eigenvalue of the adjacency matrix ). **B** Whole-brain regression of the negative of the successor representation onto neural activity, with , reveals linear scaling with the successor representation in the hippocampal-entorhinal formation (family-wise error-corrected at peak level within a bilateral entorhinal cortex/subiculum mask, left p = 0.005, peak t22= 4.80 [-18, -19, -25] and right p = 0.007, peak t22 = 4.66, [24, -22, -25]. Both clusters also survived SVC for a larger ROI comprising the hippocampus, parahippocampal cortex and entorhinal cortex, left p = 0.02 and right p = 0.03, see ROIs in Figure 2 – Figure supplement 1B). The statistical maps are thresholded at p < 0.01 for visualisation.