An abstract relational map emerges in the human medial prefrontal cortex with consolidation

Alon Baram[1], Hamed Nili[1,2], Ines Barreiros[1], Veronika Samborska[1], Timothy E. J. Behrens*[1,3] and Mona M. Garvert*[1,4]

[1]Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

[2]Department of Excellence for Neural Information Processing, Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

[3]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, UCL, London, UK

[4]Juniorprofessorship of Neuroscience, Faculty of Human Sciences, Julius-Maximilians-University of Würzburg, Germany

* Equal contribution

Corresponding authors: alon.baram@ndcn.ox.ac.uk, mona.garvert@uni-wuerzburg.de

Abstract

Understanding the structure of a problem, such as the relationships between stimuli, supports fast learning and flexible reasoning. Recent theoretical suggestions have highlighted the usefulness of explicit structural representations that are fully divorced from sensory details for generalisation. Rodent work has suggested that abstraction of structure occurs gradually, over time, in cortex. However, direct evidence of such explicit relational representations in humans is scarce, and its relationship to consolidation mechanisms is underexplored. Here, we use a graph-learning paradigm to find such a relational map in the human medial prefrontal cortex. Importantly, this representation was absent early after learning but emerged on the time scale of days. These results shed new light on neural representations underlying the remarkable human ability to draw accurate inferences from little data.

Introduction

Humans and other animals constantly find themselves in situations they have never experienced before. How does our brain manage to choose adaptive behaviors given this diversity of experiences? Luckily, our world is replete with statistical structure, and similar relationships hold across related experiences (Shepard, 1987). If the brain effectively extracts and organizes information about the statistical relationships between experiences - forming what is known as a “cognitive map” (Behrens et al., 2018; Tolman, 1948) - it can generalize knowledge from one experience to similar situations.

While cognitive maps may initially form for specific stimuli or environments, they can also support more abstract forms of learning. This allows for the generalization of knowledge across different situations that follow a similar structure. This type of abstraction has been described in various ways: as a “schema” in human behaviour (Bartlett, 1932; Piaget, 1948) and memory research (Ghosh and Gilboa, 2014; Tse et al., 2007, 2011); as “learning set” in animal learning theory (Harlow, 1949); and as “transfer learning” (Bozinovski and Fulgosi, 1976) and “meta-learning” (Wang et al., 2018) in machine learning.

One particularly effective way of achieving this abstraction is by separating (or “factorizing”) the representation of the relational structure of the environment from its specific sensory details (Behrens et al., 2018; Whittington et al., 2020). A potential locus for such an abstracted representation is the medial prefrontal cortex (mPFC), which has been implicated in recognising commonalities across experiences in the schema literature (Baldassano et al., 2018; El-Gaby et al., 2023; Gilboa and Marlatte, 2017; van Kesteren et al., 2013; Tse et al., 2007, 2011) and using these for inference (Zeithamova and Preston, 2010), as well as in concept learning (Kumaran et al., 2009; Mack et al., 2020), structural generalisation of reinforcement learning problems (Baram et al., 2021; Samborska et al., 2022) and generalisation of spatial tasks across different paths (Kaefer et al., 2020; Morrissey et al., 2017; Tang et al., 2023; Yu et al., 2018).

In addition, a recent line of research, inspired by celebrated findings from rodent studies in spatial cognition, has shed new light on the relational and structural properties

of mPFC codes. Like the entorhinal cortex, the human mPFC contains grid cells (Jacobs et al., 2013), which are active on a regular triangular lattice during spatial navigation (Hafting et al., 2005). Importantly, grid cells maintain their cell-to-cell activity covariance structure across different spatial environments (Barry et al., 2012; Fyhn et al., 2007; Yoon et al., 2013), suggesting they encode the statistical relationships common to all spatial environments, irrespective of specific sensory details. Complementary theoretical work suggests grid cells represent an efficient compression of the transition structure of 2D environments (Banino et al., 2018; Behrens et al., 2018; Dordek et al., 2016; George et al., 2021; Mathis et al., 2015; Stachenfeld et al., 2017; Whittington et al., 2020). Such a representation is highly useful for generalisation to novel tasks with the same 2D structure.

Notably, this phenomenon extends beyond spatial environments. A flurry of studies using a technique to indirectly index grid cell activity (hexadirectional coding (Doeller et al., 2010)) suggests that the same brain network (broadly referred to as the “default mode network”) that displays “grid-like” activity during spatial navigation (Doeller et al., 2010) also displays it during navigation to novel “locations” in a wide range of non-spatial and abstract 2D tasks (Bao et al., 2019; Bongioanni et al., 2021; Constantinescu et al., 2016; Park et al., 2021; Veselic et al., 2023). In most of these studies, the strongest grid-like signals emerge in the pregenual mPFC, suggesting this region encodes the 2D relational structure in these tasks. While these studies focused on 2D environments, theory suggests the 2D grid representation is merely a special case of abstract structural encoding applicable to a much wider range of task structures (Whittington et al., 2020, 2022).

The mPFC has also been implicated in a related, yet distinct literature on the consolidation of memories. Influential theoretical ideas suggest that during consolidation (i.e. on the time scale of days or weeks) cortex abstracts structural knowledge across multiple memories (Kumaran et al., 2016; McClelland et al., 1995), with neural replay (Foster, 2017) playing a key role in the process. Recent additions to this theory suggest mechanisms by which replay can be tuned to balance the trade-off between memorisation and generalisation (Sun et al., 2023). The mPFC has emerged as an important area for

this process of abstraction over multiple days. For example, mPFC activity in rats initially reflects specific perceptual features of learnt associations, and only later encodes generalisable relational features that are shared between associations (Morrissey et al., 2017; Richards et al., 2014; Tse et al., 2011). However, the extent to which similar dynamics apply in humans remains unknown (Tompary and Davachi, 2017), and a more precise understanding of the form of such abstract representations has remained elusive.

Here, we hypothesized that the mPFC abstracts an explicit relational representation of the environment over the course of multiple days. We used fMRI adaptation (Barron et al., 2016; Grill-Spector and Malach, 2001) and representational similarity analysis (RSA) (Kriegeskorte et al., 2008; Nili et al., 2014) to assess whether such an abstract map emerges after training participants extensively on two graph structures with the same underlying structure. Graphs were defined by the associations between objects. After training, we tested the representation of the two graphs in two scanning sessions separated in time by up to one week. The objects and the graphs were identical in both contexts, but pairwise distances between objects were decorrelated through the distribution of the objects on the two graphs. This allowed us to simultaneously test where in the human brain the currently task-relevant map and the currently task-irrelevant map are represented. Crucially, we were also able to test for an abstracted map representation which reflects the underlying graph structure, irrespective of the specific stimuli associated with a node.

We found that the medial temporal lobe represented the relevant and irrelevant graphs from the first scanning session, whereas mPFC represented this information in the later scanning session. Importantly, in addition to the two specific graphs, a representation of the relational structure that was independent of the specific stimulus identities emerged in the mPFC over time. This was corroborated in two fully independent analyses, demonstrating that mPFC slowly extracted an abstracted map that could presumably be rapidly transferred to facilitate learning in new, but similarly structured situations.

Results

Subjects learn and flexibly switch between two cognitive maps

We designed a task to examine how the representation of abstract cognitive maps changes over on the timescale of days. 23 participants were extensively trained on the relationships between a set of objects, each forming a graph (Figure 1A, B). Importantly, the two graphs shared the same non-symmetric structure and differed only in terms of the distribution of objects across the graph´s nodes. Each object´s position on the blue graph thus corresponded to a unique object in the same position on the red graph. The objects were distributed such that pairwise distances between the objects were uncorrelated across the two graphs.

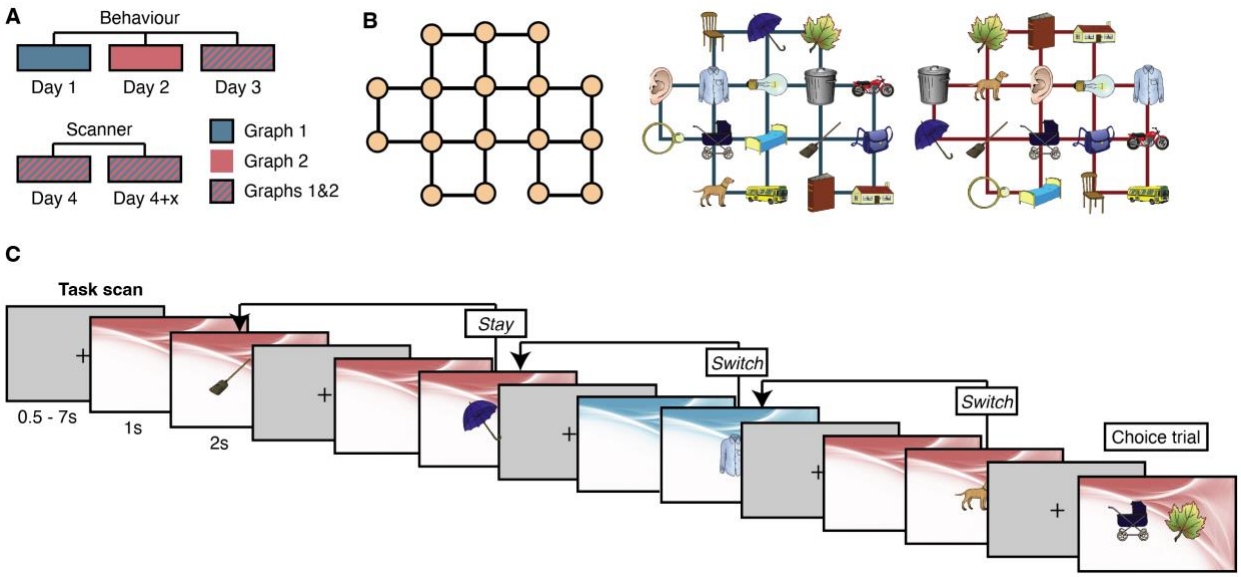

Figure 1. Experimental design. A Experimental design. Participants were trained online on the two graph structures over three subsequent days (Supplementary Figure 1). On days 1 and 2, only one of the two structures was trained, on day three both structures were trained. Scan sessions took place on day 4 (session 1) and between one day and one week after the first session (session 2). **B** Graph structure (left) and example object distribution for the two contexts (middle and right). Across subjects, object distributions were randomized, and pairwise distances between objects on the two graphs were orthogonalized for every subject. **C** Task in the scanner. Participants were exposed to random sequences of objects after a brief context primer. Trials could be classified as *stay* trials (no change in context from trial to trial) and *switch* trials (change in context from one trial to the next). Occasionally, two objects were presented simultaneously. Participants were asked to indicate which of the two presented objects could be reached from the previously presented object in fewer steps in the relevant context (*choice* trials).

Participants performed different tasks online, aimed at teaching them the pairwise associations between objects as well as sequences through the graph (Supplementary

Figure 1). The underlying structure of the graph was not made explicit. Each graph was learned separately on two subsequent days to minimize interference. On a third day of training, participants repeated most of the tasks for both contexts to ensure that they learned to rapidly switch between graphs in a context-dependent manner. In all tasks, the relevant graph was indicated by the context, i.e. a red or blue background image (order counterbalanced across subjects).

After three subsequent days of learning the object relationships, we presented subjects with sequences of objects in the scanner (Figure 1C). The first scanning session took place on the day after participants completed the training (day 4). The second scanning session took place between one day and one week after the first scanning session. The task in the scanner involved frequent context switches. The relevant context was always signaled by the same background color that subjects had learned to associate with a particular graph's object distribution during the training procedure. On 12.5% of trials (once after each object), participants were asked to indicate by a button press which one of two presented objects could be reached in fewer steps from a preceding object on the currently relevant graph (*choice* trials). The purpose of this cover task was to keep participants´ attention focused on the embedding of each presented object within its graph structure.

Participants learned the graph structures very well, as demonstrated in a subset of participants (n = 19), who were asked to arrange both sets of objects relative to one another in an arena after the last scanning session (Figure 2A-C). The pairwise distances in this arrangement correlated strongly with the 'true' link distance matrix between pairs of objects (Figure 2D, $t_{18} = 6.64$, $p < 0.001$ for Graph 1, and $t_{18} = 6.56$, $p < 0.001$ for Graph 2, based on Fisher-transformed correlation coefficients). Moreover, the correlation strengths between the two graph structures correlated with each other ($r = 0.57$, $p = 0.01$, based on Fisher-transformed correlation coefficients) and were not significantly different from each other ($t_{18} = 0.25$, $p = 0.81$, based on Fisher-transformed correlation coefficients). These results suggest that participants encoded both graph structures equally well. Indeed, we could recover both graph structures by applying multi-dimensional scaling to the distance matrix averaged across subjects (Figure 2B-C).

More importantly, participants also used their knowledge about the graph structures to solve the task in the scanner. Performance in the choice task was significantly above chance in both sessions (both $p < 0.001$), with an improvement in performance from scanning session 1 to 2 (Figure 2E, $t_{21} = 2.09$, $p = 0.049$). Across the two sessions performance in the choice task was highly correlated with performance in the post-scan arena task ($r = 0.67$, $p = 0.002$), suggesting that participants who had learned the graph structures particularly well benefitted the most from this knowledge in the scanner.

In a logistic mixed-effects model, with subject included as a random effect, the likelihood of choosing the left or right option was significantly influenced by the difference in distances between the target object and the objects on either side. The difference in distances on the relevant graph between the target object and the chosen object had a strong positive effect on the likelihood of choosing a given option (Coef = 0.747, SE = 0.031, $z = 24.121$, $p < 0.001$, Figure 2G). This suggests that participants were more likely to choose the option on the side that was closer to the target object on the relevant graph. A smaller, yet significant, effect was also observed for the difference in distances on the irrelevant graph between the target object and the two options (Coef = 0.067, SE = 0.024, $z = 2.785$, $p = 0.005$). This indicates that the irrelevant graph also had a measurable, though smaller, influence on decision-making compared to the relevant graph

In a mixed-effects model with subject included as a random effect, the distance between the target object and the chosen object on the relevant graph had a strong positive effect on response time (Coef = 3.602, SE = 0.126, $z = 28.523$, $p < 0.001$). Additionally, the distance between the target object and the unchosen object on the relevant graph also had a significant positive effect on response time (Coef = 0.860, SE = 0.133, $z = 6.444$, $p < 0.001$). These results indicate that the length of mental simulation on the relevant graph, both from the chosen and unchosen options to the target, contributed significantly to longer response times. There was also a modest positive effect of the distance between the target object and the chosen object on the irrelevant graph (Coef = 0.248, SE = 0.118, $z = 2.090$, $p = 0.037$), suggesting some influence of irrelevant graph knowledge on response times. However, the distance between the target object

and the unchosen object on the irrelevant graph did not significantly affect response times (Coef = -0.112, SE = 0.120, z = -0.939, p = 0.348).

Next, we examined whether these effects of graph distances on response times in the scanner correlated with explicit knowledge of the graphs. We ran a separate regression for each participant (predicting, again, response times with the chosen and unchosen distances on the relevant and irrelevant graphs), and correlated across subjects the resulting parameter estimates with an index of performance in the arena task - the correlation coefficients between the arena distances and “true” distances. We found that both the distance to the chosen object and the distance to the unchosen object on the relevant graph had a stronger influence on response times for participants who performed better in the arena task (chosen: r = 0.61, p = 0.006; unchosen: r = 0.48, p = 0.04, Figure 2I, K). This suggests that participants with a better understanding of the graph structures used this knowledge more effectively to solve the choice task.

Notably, the effect of distance to the chosen object on the irrelevant graph was negatively correlated with arena task performance (r = -0.62, p = 0.005, Figure 2J), indicating that participants who knew the graph structures well were less distracted by irrelevant information. We found no significant correlation between performance in the arena task and the effect of unchosen distance on the irrelevant graph on response times, r = -0.33, p = 0.16, Figure 2L).

Together, these behavioral results suggest that participants had acquired and used structural knowledge about the relationships between objects to solve the cover task.

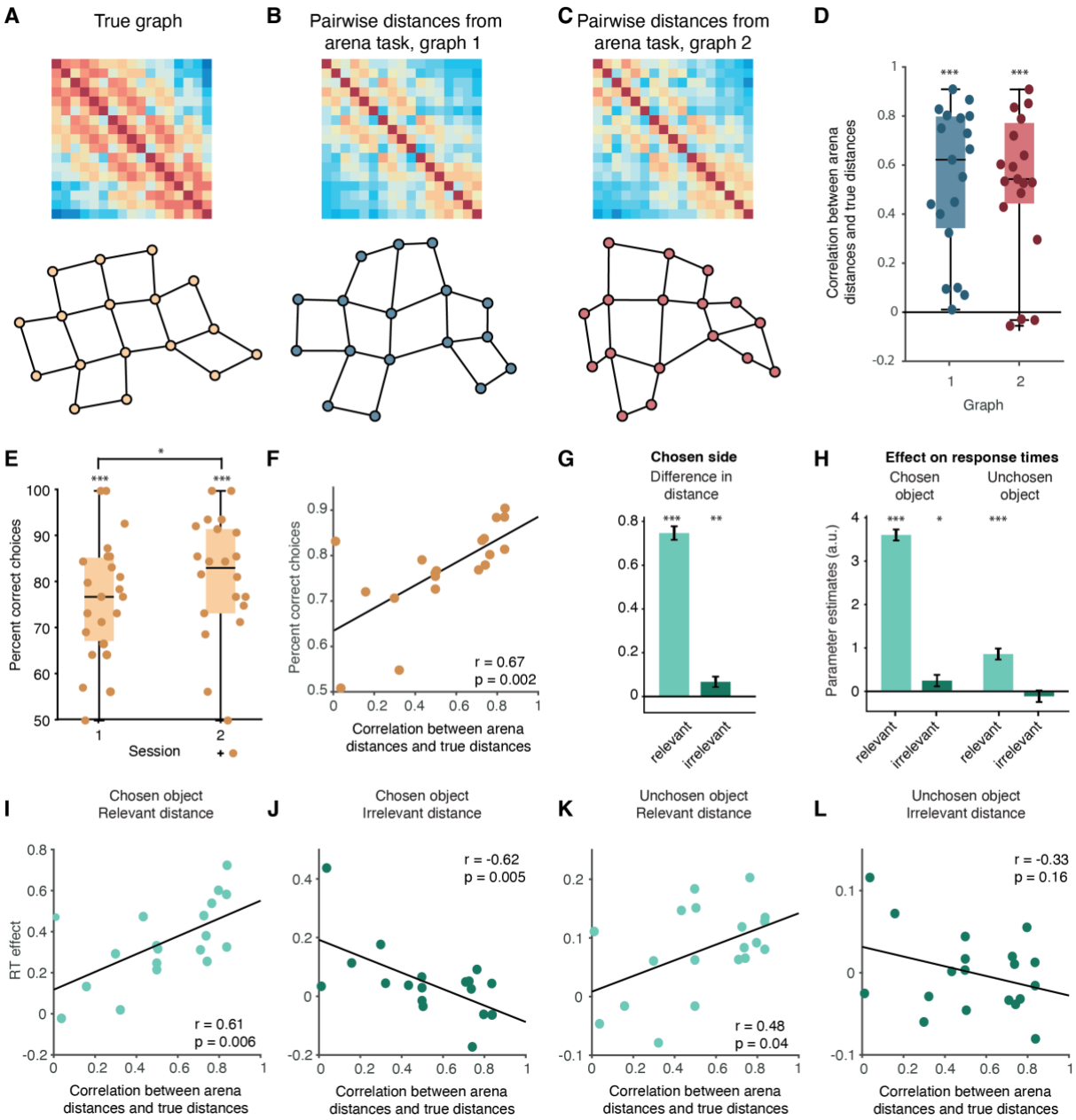

Figure 2. Behavioural performance. A Euclidian distances between object pairs (top) and reconstructed graph after projecting link distances into a two-dimensional space using multi-dimensional scaling (bottom). **B** Pairwise distances between objects of graph 1 as arranged by participants in an arena at the end of session 2, averaged across participants (top). Multi-dimensional scaling of the distance matrix was used to recover the graph (bottom). **C** Same as **B**, for graph 2. **D** Correlation between link distances in **A** and distances between positions in the arena for each graph (**B** and **C**). **E** Percent correct decisions in the choice task performed in the two scan sessions. Performance improves from session 1 to session 2. **F** Correlation between performance in the arena task and percent correct choices in the scanner, averaged across the two sessions. **G** Linear mixed effects regression coefficients reflecting the influence of relevant and irrelevant distances for chosen and unchosen objects on response times. **H** Binomial mixed effects regression coefficients reflecting the influence of relevant and irrelevant distances for chosen and unchosen objects on chosen side. In **G** and **H,** data from sessions 1 and 2 are pooled. **I** Correlation between arena performance and the influence of chosen/relevant distance on response times, averaged across the two sessions. **J** Correlation between arena performance and the influence of chosen/irrelevant distance on response times, averaged across the two sessions. **K** Correlation between arena performance and the influence of unchosen/relevant distance on response times, averaged across the two sessions. **L** Correlation between arena performance and the influence of unchosen/irrelevant distance on response times, averaged across the two sessions. *p < 0.05, **p < 0.01, ***p < 0.001

Representations of both the relevant and irrelevant maps in the MTL-mPFC circuit

To solve this context-dependent task, the brain must represent the object-embedded graph that is currently relevant in each trial. Multiple studies have shown that the MTL represents such associative graphs (Mark et al., 2024; Park et al., 2020; Schapiro et al., 2013) when these are task-relevant. Importantly, we have previously shown that MTL also represents such graphs even when learnt implicitly, and when subjects were engaged in a task for which the graph structure was irrelevant (Garvert et al., 2017; Tacikowski et al., 2024). We thus hypothesised that the MTL might simultaneously represent both the currently relevant and irrelevant graphs.

To test this, we used functional magnetic resonance imaging (fMRI) repetition suppression (Grill-Spector and Malach, 2001) – a technique which relies on the assumption that the subsequent presentation of stimuli leads to a suppressed response that scales with the distance in the representational space where the stimuli are embedded (Barron et al., 2016). We reasoned that in regions encoding a map-like representation of the relevant graph, the degree of similarity in neural representation should decrease as a function of distance (number of edges) between items on this graph. This would suggest that on *stay* trials, where the context remains the same on consecutive trials, the amount of fMRI adaptation in response to a given object should scale with the distance to the immediately preceding object on the same graph (Figure 3A). On the other hand, in regions encoding a map-like representation of the irrelevant graph, the amount of fMRI adaptation in response to the same object should scale with the distance to the immediately preceding object on the other (currently irrelevant) graph (Figure 3B). Based on this reasoning, we first looked for brain regions whose fMRI responses to each object increased as a linear function of the number of links from the preceding object on both the relevant and the irrelevant graphs. Since pairwise object distances were orthogonalized across the two graphs, we were able to look at their representation simultaneously, in the same GLM. Note that in all fMRI analyses we focused only on the object presentation trials, modelling out the choice catch trials.

We modelled the onset of objects on *stay* trials for each context separately from the onsets of objects on *switch* trials. We added two parametric modulators to each object onset regressor, one corresponding to the number of links between a given object and the previously presented object on the currently relevant graph, and one corresponding to the corresponding distance on the irrelevant graph (Figure 2A). A contrast identifying both relevant and irrelevant distances on *stay* trials across both sessions identified a cluster in the left medial temporal lobe (Figure 2B and S3A). This cluster survived family wise error (FWE) small-volume correction (SVC) within a bilateral entorhinal cortex/subiculum mask (mask from (Garvert et al., 2017) Figure S1A, p = 0.03 corrected at peak level, peak t_{22} = 4.39, [-24, -22, -28]), and trended towards significance when SVC was performed in a larger MTL mask comprising bilateral hippocampus, parahippocampal cortex, and entorhinal cortex (p = 0.053 corrected at peak level, peak t_{22} = 4.94, [-26, -22, -28], mask from Figure S1B, see masks in Supplementary Figure 2).

When analyses were performed separately for the two scanning sessions, a similar cluster in the left MTL emerged as the cortical region with the strongest representation of both relevant and irrelevant distances in session 1 (Figure 3C, SVC at peak level in the larger mask of bilateral MTL (Supplementary Figure S2b), p = 0.02, peak t_{22} = 5.47, [-26, -22, -28]). We did not find any similar cluster in MTL in session 2 that survived SVC. However, within the cluster of the session 1 effect, the parameter estimates of both maps were still significant in session 2, (t_{21} = 2.53, p = 0.02, Supplementary Figure S3B), suggesting that the MTL represents both maps in both sessions.

When the analysis was restricted to session 2, a large cluster emerged in mPFC (Figure 2D, whole-brain FWE-corrected on the cluster level, p = 0.01, [6 44 -8]). We did not find any cluster in mPFC for session 1, and session 1 parameter estimates extracted from an ROI based on the session 2 cluster were not significant (t_{21} = 0.49, p = 0.62, Supplementary Figure S3C), suggesting this effect was absent in session 1. This suggests that the mPFC comes to represent relevant and irrelevant graph distances over time, possibly as a result of consolidation mechanisms occurring over multiple days (van Kesteren et al., 2013).

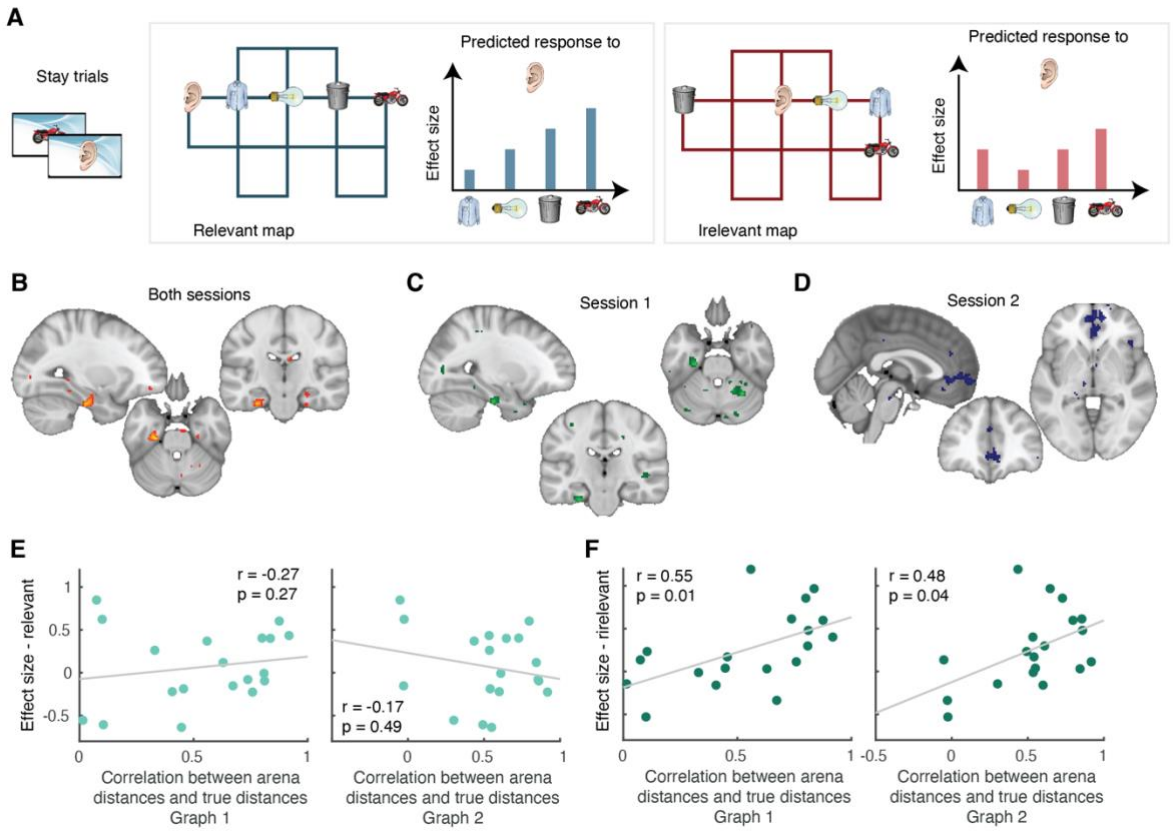

Figure 3. Relevant and irrelevant map representations. A Analysis principle. On *stay* trials, distances between an object and its predecessor in the object sequence was defined on the relevant map (here blue) and the irrelevant map (here red). Visualized are distances from the ear to four example objects. Since pairwise distances were orthogonalized for the two maps, we were able to simultaneously assess the representation of relevant and irrelevant information. **B** MTL areas representing relevant and irrelevant distances across both sessions. **C** MTL representing relevant and irrelevant distances in session 1. **D** mPFC representing relevant and irrelevant distances in session 2. **E** Correlation between performance in the arena task for graphs 1 and 2 and relevant distance parameter estimates extracted from the mPFC ROI depicted in **D**. **F** same as **D** for the irrelevant, rather than relevant distances.

Next, we tested for the behavioral relevance of these MTL-mPFC representations of the relevant and irrelevant graphs. Across subjects, we correlated the performance in the arena task with the parameter estimates in the MTL cluster (Figure 2B, parameter estimates averaged across both sessions) or the mPFC cluster (Figure 2D, parameter estimates from session 2). We found that the session 2 BOLD correlate of the irrelevant distances in mPFC predicted performance in the arena task ($r = 0.55$, $p = 0.01$ for graph 1 and $r = 0.48$, $p = 0.04$ for graph 2, Figure 3F). No such relationship could be found for the session 2 mPFC representation of relevant distances (Figure 3E) or for the MTL cluster. One interpretation of this result is that subjects who knew the graphs better might

have actively suppressed the irrelevant graph – with the stronger mPFC irrelevant graph representation being an index to this suppression.

Interestingly, visual inspection of the parameter estimates from the MTL “both maps, both sessions” cluster (Fig 3B) suggests that in session 2 there is a stronger representation of the irrelevant graph than the relevant graph, though this difference was not significant (Supplementary Figure S3A). We hypothesized that an active suppression of the irrelevant graph should be most prominent immediately after a context switch, when the recently relevant graph requires the most suppression. To explore whether this is indeed the case, we split the *stay* trials to either those immediately preceded by a *switch* trial, or those which were further away from a *switch* trial. A 2x2 ANOVA with the factors “relevance” and “distance from *switch*”, performed separately on parameter estimates from both sessions and in both the MTL and mPFC ROIs, revealed a significant interaction for the MTL cluster in session 2 ($F_{1,21} = 5.12$, $p = 0.03$, Supplementary Figure S4). This suggests that in session 2 the irrelevant graph representation in the MTL was stronger immediately after a switch – precisely when suppression of the recently relevant, but now irrelevant graph was most necessary.

Together, the results so far suggest that the MTL represents knowledge about the specific graph structures in both sessions, whereas this representation emerges in mPFC only in session 2. In both cases, relevant and irrelevant graphs are represented simultaneously, although irrelevant distances appear to be represented more strongly session 2, possibly reflecting behaviorally relevant active suppression of the irrelevant information.

A generalisable abstract map in the mPFC

The medial prefrontal cortex, where knowledge about relevant and irrelevant object-embedding graphs emerges over time, is known to abstract the commonalities across experiences and represent schematic knowledge (Baldassano et al., 2018; El-Gaby et al., 2023; Gilboa and Marlatte, 2017; Samborska et al., 2022). Moreover, in humans mPFC contains grid cells (Jacobs et al., 2013) and grid-like signals (Bao et al., 2019; Constantinescu et al., 2016; Doeller et al., 2010; Park et al., 2021; Primates: Bongioanni

et al., 2021; Veselic et al., 2023), suggested to encode statistical structure in 2D spaces in a way that reflects commonalities and ignores particularities across tasks and environments (Behrens et al., 2018; Whittington et al., 2020). Such abstract structure representations could be useful for transferring structural knowledge to new situations and thereby speed up learning.

In the context of our experiment, such structure-coding would suggest the presence of an abstract map that ignores the specific context a stimulus is presented in. To test for an abstract graph representation, we defined a cross-map distance as the distance between the position of the currently presented object on its map and the position of the preceding object on the other map on *switch* trials (Figure 4A). Such a cross-map suppression effect would only be expected in brain areas that represent structure explicitly, and separately from the stimuli themselves. Brain areas that encode the relationships between stimuli in a context-dependent manner would show the representation of the relevant or irrelevant maps, but no cross-map suppression effect.

A whole-brain analysis across sessions identified a significant cluster representing cross-map distances in mPFC (Figure 4B, FWE-corrected on the cluster level, $p = 0.003$, [2, 42, 8]). This effect is driven by an emergence of a cross-map suppression effect in session 2: No cluster survives corrections for multiple comparisons when the analysis is restricted to session 1, whereas a strong cross-map suppression effect can be observed in mPFC in session 2 (Figure 4C, FWE-corrected on the cluster level, $p < 0.001$, [2, 44, 8]). This is confirmed in an independent analysis where parameter estimates are extracted directly from a region-of-interest defined based on the existence of a relevant and irrelevant graph representation in *stay* trials in session 2 (Figure 4D). Within this ROI we find a significant cross-map suppression effect in session 2 ($t_{21} = 2.1$, $p = 0.048$), but not in session 1 ($t_{21} = -0.2$, $p = 0.85$). While these indirect repetition suppression effects do not allow us to make any inference about the particular types of cells involved in this representation, it is notable that these cross-map suppression effects are stronger in those parts of mPFC that were shown to display grid-like coding of abstract relationships in session 2 ($t_{21} = 2.86$, $p = 0.01$) but not in session 1 ($t_{21} = 0.09$, $p = 0.93$, difference between session 2 and session 1: $t_{21} = 2.38$, $p = 0.03$, Figure 4E, (Constantinescu et al.,

2016)).

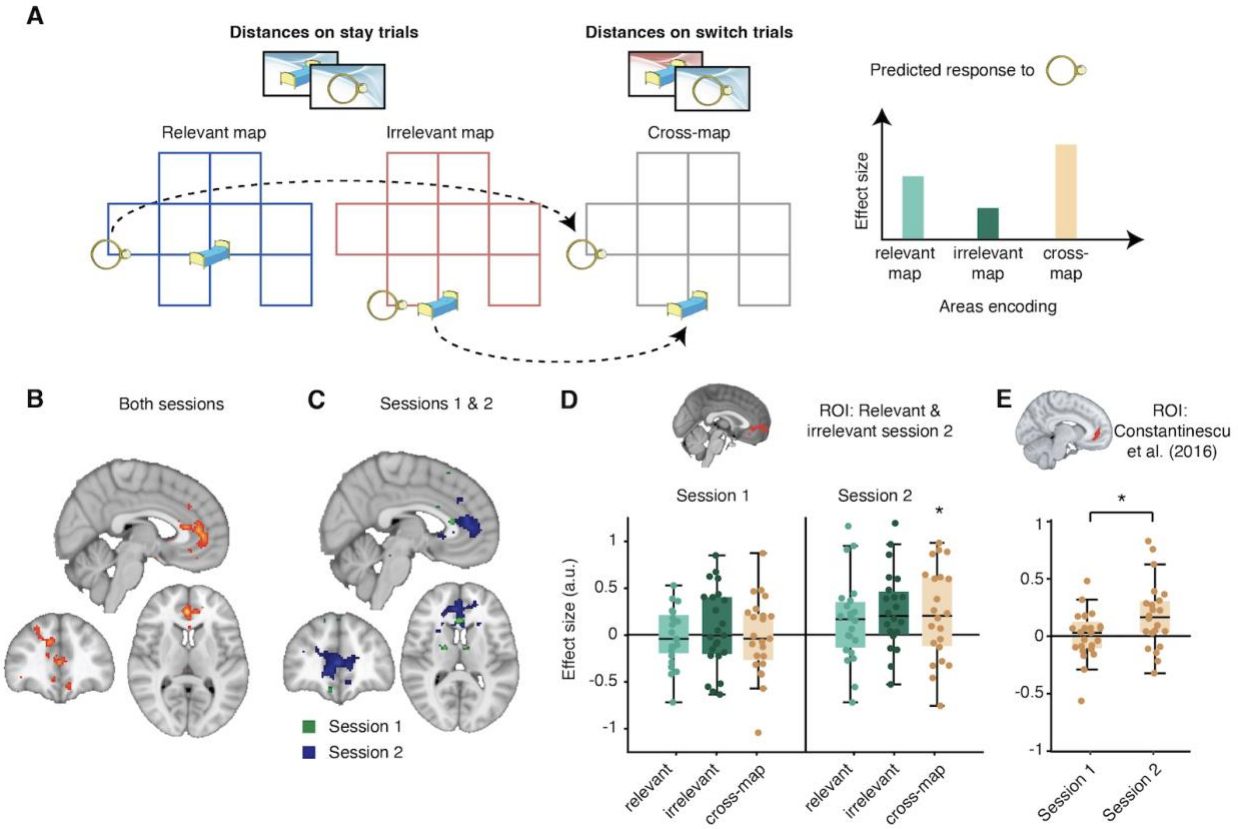

Figure 4. Abstract structure knowledge emerges in mPFC over time. A Logic of the cross-map suppression analysis on *switch* trials. Cross-map distances are defined as the distances between the position of the currently presented object on its map and the position of the preceding object on the other map. This measure corresponds to distances expected in an area that ignores the specific stimuli in a position, and instead reflects an abstract relational structure. **B** Across both sessions, cross-map suppression is found in in the medial prefrontal cortex. **C** A cross-map suppression only emerges in session 2. The session 1 effect is overlayed on the session 2 effect. **D** A cross-map suppression effect is also significant in session 2 in the ROI representing relevant and irrelevant distances in session 2. Effect sizes for relevant and irrelevant distances are displayed for visualization only. **E** A cross-map suppression effect is also significantly stronger in session 2 in an ROI defined based on grid-like coding of abstract relationships (Constantinescu et al. 2016). *$p < 0.05$

If the mPFC reflects abstract structure knowledge rather than two separate graphs, then such a representation would additionally predict the presence of a position code, where objects that share the same position on the two graphs have a more similar representation to each other than to all other objects. Since position repeats on *switch* trials were very infrequent, this analysis cannot be performed using repetition suppression. Instead, we used representational similarity analysis (RSA), which enables us to compare the similarity of neural representations across voxels between conditions, using data from all non-choice trials (Figure 6A, Kriegeskorte et al., 2008). We performed

this analysis on the cortical surface to ensure that only adjacent cortical voxels are included together in the same searchlight.

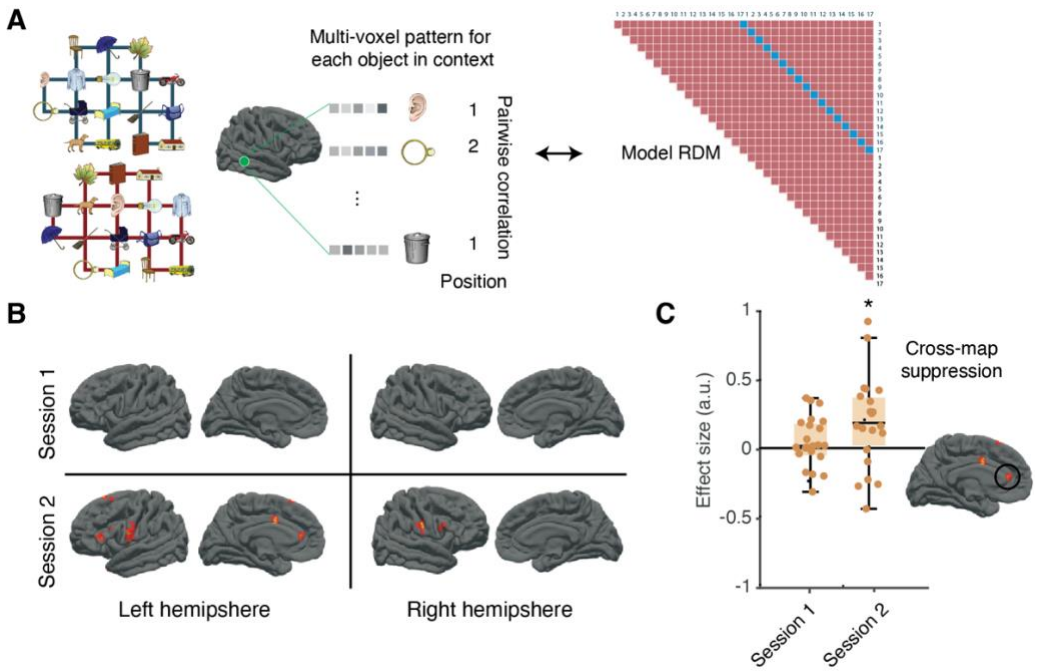

Figure 5. Position code emerges in the same region that shows a cross-map suppression effect. A Representational similarity analysis. We extracted the dissimilarity of patterns across voxels separately for each of the 17 objects on each graph and each context. We predicted the emergence of a position code, i.e. higher similarity for two objects that occupied the same position on the graph (e.g. blue ear and red bin) compared to all other objects. **B** No position code could be observed in session 1. In session 2, various clusters emerged, including one in pregenual mPFC. **C** In the mPFC cluster displaying a position code in session 2, also a cross-map suppression effect emerges in session 2. *p<0.05

In session 1, a whole-cortex searchlight analysis revealed no position code in any part of the brain. However, in session 2, various prefrontal clusters emerged, including an mPFC cluster that overlaps with the region where we found a cross-map suppression effect in session 2 (Figure 6B). To directly confirm that a cross-map suppression effect and a position code emerged in the same region of mPFC, we extracted parameter estimates for the cross-map suppression effect from a ROI defined based on a significant position code transformed back into volumetric MNI space (Figure 6C). We found that indeed in this region, no cross-map suppression effect was present in session 1 (t_{22} = 0.92, p = 0.37), but an effect emerged in session 2 (t_{21} = 2.65, p = 0.02). Notably, the fMRI suppression-based cross-map suppression effect and the multivariate position code measure very different aspects of the BOLD signal (Barron et al., 2016). It is striking that

both effects emerge in the same part of the mPFC, and in both cases only in session 2. Taken together, these converging results point to an abstract representation of task structure that emerges in mPFC in session 2, following consolidation.

Discussion

Here, we used a graph-learning paradigm to investigate how abstract relational structures are represented in the brain and how these representations change over time. Participants learned two different graphs characterized by the same underlying associative structures, but a different distribution of stimuli. The retrieval of the graphs initially depended on the MTL, but over the course of several days became dependent on the prelimbic region of the mPFC (Takehara-Nishiuchi and McNaughton, 2008; Takehara et al., 2003). By that point in time, an additional cross-map suppression effect as well as a position code emerged in the mPFC, pointing to the emergence of an abstract, stimulus-independent structural representation that was absent in session 1. Separating the representation of task structure from stimuli is useful, because it facilitates the generalisation of knowledge across environments that share the same statistics (Behrens et al., 2018; Mark et al., 2024; Whittington et al., 2020). If task knowledge is organised in such a factorised form, then solutions to new tasks do not have to be learnt afresh, they can instead be inferred (Behrens et al., 2018; Mark et al., 2020).

While the mPFC has been implicated in many human studies probing schematic and structural encoding (Baldassano et al., 2018; Baram et al., 2021; Van Kesteren et al., 2010) as well as in the human cognitive maps literature (Park et al., 2020; Schapiro et al., 2013), to our knowledge this is the first time the changes in such an explicit abstract map following consolidation have been investigated. Our results suggest that knowledge about the graph structures continues to be processed after the first scan even in the absence of further training and becomes more abstract over time. This is in line with theories of systems consolidation positing that memories change qualitatively after acquisition, from an encoding of specific episodic information to the abstraction of statistical structure that is shared between episodes (Richards et al., 2014). In this way, the neocortex gradually

builds semantic knowledge of the world in a process that can take many weeks (McClelland et al., 1995; O'Reilly and McClelland, 1994). In rodents, such consolidation effects in the mPFC manifest themselves in a gradual loss of selectivity for perceptual features characterizing specific stimuli and an increase in the representation of relational features that are shared between stimuli (Morrissey et al., 2017). In humans, featural overlap between memories was represented in mPFC (as well as hippocampus) one week after, but not immediately after encoding (Tompary and Davachi, 2017).

Indeed, memory reactivation during sleep plays a critical role in extracting rules and identifying abstract patterns across experiences (Robertson, 2018; Stickgold and Walker, 2013). A potential mechanism that might be involved in this abstraction is neural replay (Foster, 2017), which in humans reorganises experiences according to pre-learnt structures (Liu et al., 2019). Replay also contains abstract structural codes in addition to the sensory ones (Liu et al., 2019) and has been shown to be constrained by the structure of the task in a way that guides relational inference (Schwartenbeck et al., 2023). Recent work suggests that the amount of replay of specific experiences depends on the reliability versus generalisability of the environment, thus balancing the trade-off between memory and generalisation (Sun et al., 2023). Future work might explore a direct relationship between replay and the structural mPFC representations such as those we find here.

It is worth noting that while the session 2 mPFC representation of the relevant and irrelevant maps on *stay* trials can be interpreted as the result of the same abstract representation we find on *switch* trials, we cannot rule out the possibility that the mPFC additionally represents the specific maps and their sensory details. This is reminiscent of recent findings in rodents (El-Gaby et al., 2023) where mPFC cells implemented an algorithm that instantaneously represents the entire sequence of future behavioural steps (and hence bound to the particularities of the current task at hand) in a manner that was flexible and reconfigurable between many tasks with the same structure (and hence inherently relational and abstract).

While not the focus of the current manuscript, it is worth noting the rapid context-switching task subjects experienced in the scanner can be related to other proposed

functions of the prefrontal cortex, such as cognitive control (Miller and Cohen, 2001). Indeed, the ventromedial PFC contain rich representations of context and its associated value of both relevant and irrelevant contexts (Moneta et al., 2023). However, we do not believe our effects of cross-map and position representations can be explained as cognitive control signals.

Another possible view on such context-switching tasks is through the lens of spatial cognition. In remapping experiments with rapid context switching ("teleporting"), the hippocampal representation rapidly flickers between the representations of both contexts, effectively implementing an inference process of the current context (Jezek et al., 2011). While our manuscript does not focus on the hippocampus, we note that our MTL effect (partly including hippocampus) of the relevant and irrelevant maps is compatible with such a remapping effect, as a representation of a relevant map (which is strong particularly in session 1, see Figure S3) is indeed what is expected from hippocampal remapping.

Methods

Participants

Twenty-three volunteers (aged 18 – 32 years, mean age ± standard deviation 23.2 ± 3.8 years, 14 female) participated in this study. One participant did not return to the lab for the second scan and only the first scan session of this participant was included in the analyses. All subjects reported normal or corrected-to-normal vision and no history of neurological or psychiatric disease. Participants gave written informed consent and the study was approved by the Medical Sciences Inter-Divisional Research Ethics Committee (IDREC) of the University of Oxford (ref. number R49191/RE001). Participants were naïve to the purpose of the experiment. The scans took place at the Oxford Centre for Functional MRI of the Brain (FMRIB), Wellcome Centre for Integrative Neuroimaging.

Experimental structure

The experiment took place over three training and two scan sessions (Figure 1A). The three training sessions and the first scan session took place over four subsequent days. The second scanning session was scheduled between one and seven days after the first scan. In the training sessions participants performed a set of tasks to learn the relational structures. Participants were free to do the training in their own time at home through the online testing platform Prolific (www.prolific.com). These tasks were run in custom-written JavaScript, CSS and HTML scripts. Because of problems with data storage on Prolific, for some subjects a few task blocks during the pre-training could not be recorded. However, all participants completed all tasks.

Stimuli

Thirty-one coloured and shaded object images that were similar in terms of 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 17 objects was randomly chosen and assigned to the 17 nodes of the two graph structures (Figure 1B). The positions of the objects on the two graphs were randomized such that pairwise distances between objects were orthogonal between graphs. The identity of the currently relevant graph was indicated by a blue or red background image. The colour associated with the graph learnt on the first day of training was randomly assigned and the alternative color was used for the second graph. Graphs one and two were initially learned separately on days one and two, and jointly on day three before the first scan.

Pre-training tasks

On each day, participants performed a set of computer-based tasks to learn the graph structures. In most of the tasks, participants could gain or lose coins depending on their performance. At the end of the experiment, the accumulated coins were converted to £ (every 100 coins = 1£ extra) and this money was paid in addition to the baseline payment of £7 per behavioral session. Visual feedback after each trial indicated whether a response was correct. On all days of the experiment, subjects were shown the number of correct trials and/or the number of coins collected at the end of each task.

To participate on subsequent days of the study, subjects were required to meet a minimum performance criterion of 400 coins collected on the first day of the experiment. This corresponds to about 70% of the total number of coins that could be obtained in the session. All trials and breaks between tasks were self-paced.

Tasks

The set of tasks was designed to teach participants the associations between pairs of objects and to familiarize them implicitly or explicitly with the object transitions in the graph (Supplementary Figure 1A).

Task 1 - Object pair learning

Subjects were presented twice with each one of the 24 object pairs (i.e. objects directly connected to each other by a link in the graph structure) and asked to memorize each pair (Supplementary Figure 1B).

Task 2 - Associative memory task

Subjects were presented twice with each object pair as well as with 48 pairs of objects they did not learn to associate with each other in a random order and asked whether the two objects were paired or not (Supplementary Figure 1C). Subjects gained or lost a coin for each correct or incorrect response, respectively.

Task 3 - Object orientation task

Subjects were exposed to object sequences generated from a random walk along the graph structure, such that each object would only be followed by a neighboring object on the graph structure (Garvert et al., 2017). Objects were presented in one of two orientations, which were mirror-images of each other (Supplementary Figure S1D). Subjects had to learn which orientation was associated with which button press. Thus, in this task the graphs were learnt in an implicit, rather than explicit manner. Subjects would gain or lose a coin for each correct or incorrect response, respectively. The task consisted of 100 trials.

Task 4 - Three-alternative forced choice task

Subjects were presented with one object at the top of the screen and had to pick the object it was paired with from three objects presented below (Supplementary Figure S1E). Each object was presented as many times as the number of objects it was linked to on the graph, such that each association was tested once. Subjects would gain or lose a coin for each correct or incorrect response, respectively.

Task 5 - Directed navigation task

In this task subjects navigated through the graph structure to find a given target object. The target object was displayed at the top of the screen. In the center of the screen the object corresponding to the current state were presented. Above this object, two cards were presented which were paired with this object. When choosing the next object, the object at the center was replaced by the selected object and the options were replaced by two new objects the chosen object was paired with (Supplementary Figure S1F). This means that by clicking through the objects, subjects moved from association to association until they reached the target. Below the objects, a progress bar and a number of coins were presented. At the start of a trial, the number of coins corresponded to the distance on the graph between the starting and the target object. Selecting the object option closer to the target would move the bar by the distance corresponding to (the total length of the bar/minimum number of steps to reach target) to the right. Conversely, incorrect choices would move the progress bar to the left and a coin was subtracted from the amount that could be earnt on the trial. When subjects filled the progress bar, they got the number of coins displayed below. This means that the fewer steps it took a subject to reach the target, the more coins they gained. At the end of each trial, subjects were presented with the sequence of objects they had chosen to encourage learning. This task consisted of 30 trials.

Task 6 - Distance difference task

On each trial, participants were presented with a target object at the bottom of the screen and two goal objects presented on top of the target object. Participants were

required to select the option that they would reach in fewer steps if they moved from association to association like in the directed navigation task from the target object (Supplementary Figure S1G). In other words, participants were required to select the option closer to the target object on the graph. Participants would gain or lose a coin for each correct or incorrect response, respectively. This task consisted of 30 trials.

fMRI experiment

Before the scanning session on the next day, participants performed a brief reminder training session consisting of one block of task 4 (three-alternative forced choice task) and one block of task 6 (goal navigation task). They were also familiarized with the task they would perform in the scanner through a brief pre-scan training session that was the same as the task performed in the scanner (15 trials).

In the scanner, subjects were presented with random object sequences. Each of the 17 objects was presented exactly four times per context in each experimental block, resulting in a total number of 136 presented objects per block. 1 second before stimulus onset, the background colour changed according to the relevant graph. Stimuli were then presented with the stimulus in context for 2 seconds, followed by a jittered inter-trial interval (ITI) generated from an exponential distribution with a mean of 1.7 seconds and truncated between 0.5 and 7.5 seconds (Figure 1C).

While observing the object sequences subjects performed a cover task of infrequently reporting by button press which one of two presented objects they would reach in fewer steps. This choice task occurred after random intervals exactly once after each object, i.e. in 12.5% of the total number of trials. Each correct button press was rewarded with £0.10 paid out in addition to the baseline fee to ensure that subjects attended to the stimuli. Subjects received a brief training on this task before they performed it in the scanner.

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 were acquired using a 32-

channel head coil on a 3 Tesla Magnetom Prisma scanner (Siemens, Erlangen, Germany). A T2*-weighted echo-planar imaging (EPI) sequence with a multi-band acceleration factor of 3 and within-plane acceleration factor (iPAT) of 2 was used to collect 72 transverse slices (interleaved order) with TR = 1.235 s, TE = 20 ms, flip angle = 65°, voxel resolution = 2x2x2 mm. Slices were tilted by 30˚ relative to the axial axisThe first five volumes of each block were discarded to allow for scanner equilibration. After two fMRI blocks, a T1-weighted anatomical scan with 1x1x1 mm resolution was acquired. In addition, a whole-brain field map with dual echo-time images (TE1 = 4.92 ms, TE2 = 7.38 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. Contrast images were later spatially normalized by warping subject-specific images to an MNI (Montreal Neurological Institute) reference brain, and smoothed using a 6-mm full-width at half maximum Gaussian kernel. All pre-processing steps were performed with FSL. Note that for the RSA analyses we used unsmoothed preprocessed data, and later smoothed the RDMs.

Behavioural data analysis

All data analysis of the training sessions, as well as of performance in the arena task after session 2, was performed using custom-written Matlab scripts (Mathworks, USA).

To obtain a metric for performance in the arena task, we first organized each subject's responses into pairwise distance matrices for each pair of objects. For visualisation purposes, we projected the matrices to a two-dimensional space using multi-dimensional scaling (MDS, Figure 2A-C bottom). Next, we correlated these distances with the "true" distance matrix, based on the number of shortest paths between object pairs (Figure 2D), as well as correlating the matrices of the two graphs with each other. To test for significance of these correlations across subjects we ran a one-sided t-test on the Fisher-transformed correlation coefficients.

Analysis of the behavior in the choice trials during the scans was performed using Python. We first ran a logistic mixed-effects model using Pymer4's Lmer function, with subject included as a random effect, where the dependent variable was the likelihood of choosing the left or right option and the independent variables were the differences in distances between the target object and the objects on either side (Figure 2G).

We then ran a mixed-effects linear model using Statsmodels' MixedLM function, with subject included as a random effect, where the dependent variable was the response times in the choice trials and the four independent variables were the distances between the target object and the 1) chosen object on the relevant graph; 2) unchosen object on the relevant graph; 3) chosen object on the irrelevant graph; 4) unchosen object on the irrelevant graph (Figure 2H). Next, we wanted to obtain subject-by-subject scores of the effects from this analysis, in order test for a correlation with the performance in the arena task (Figure 2I-L). We achieved this by running a separate linear regression for each subject, with the same dependent and independent variables as above.

fMRI data analysis

We modeled the onsets of objects separately for each context and for *stay* and *switch* trials. Each regressor additionally contained two parametric modulators, one reflecting distance between the current object and the preceding object on the relevant map and one reflecting the distance between the current object and the preceding object on the irrelevant map.

Contrasts were defined as relevant + irrelevant distances on stay trials or as relevant distances on *switch* trials, respectively. The contrast images of all subjects from the first level were analyzed as a second-level random effects analysis. We report our results in the hippocampal–entorhinal formation, as this was our a priori ROI for the representations of the relevant and the irrelevant maps, at a cluster-defining statistical threshold of $p<0.01$ uncorrected, combined with 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 (Chadwick et al., 2015), (Supplemental Figure 2A). The second

mask also contained other medial temporal lobe regions implicated in encoding physical space and comprised the hippocampus, entorhinal cortex, and parahippocampal cortex, as defined according to the maximum probability tissue labels provided by Neuromorphometrics, Inc. (Supplemental Figure 2B). 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 2A at an uncorrected threshold of $p<0.01$ for completeness. While we used masks to correct for multiple comparisons in our ROI, all statistical parametric maps presented in the manuscript are unmasked. All the clusters used in the manuscripts for extraction of (independent) parameter estimates were obtained by thresholding at $t=2.5$.

Representational similarity analysis

Representational similarity analysis (Kriegeskorte et al., 2008) was performed on the surface in subject space. We first ran a GLM where we modelled the presentation of each object in one of the graphs as a different condition (without any distinction between *stay* or *switch* trials). We used Freesurfer on the T1 scans to create a gray matter surface map and then used a searchlight procedure (Kriegeskorte et al., 2006) to compute representational dissimilarities for each coordinate on the surface. The searchlight size was defined such that each coordinate reflected the representational dissimilarity for up to 100 voxels within a 10 mm radius around it. Representational dissimilarity was computed as the Euclidian distances between voxel patterns after spatially whitening the data within the searchlight (i.e. the Mahalanobis distance, (Walther et al., 2016)). This was computed separately for each subject, session and block. To obtain one RDM per subject and session, RDMs were averaged across the four blocks. The average neural RDM was then related to each model RDM using Kendall's tau-a (Nili et al., 2014)

Each whole-brain map coordinate [x,y] thus contained the correlation of the brain's dissimilarity matrix with a model RDM when that coordinate is in the centre of the searchlight. To compute group statistics, we moved each map into the fsaverage space, smoothed the maps and ran PALM (Permutation Analysis of Linear Models), a tool that

allows inference on the surface using permutation methods (Winkler et al., 2014). Results are considered significant if they survive FWE-correction at $p = 0.05$.

Model RDMs were a 34x34 matrix of ones, with zeros at positions that corresponded to the same position.

Acknowledgements

T.B. is supported by a Wellcome Principal Research Fellowship (219525/Z/19/Z), a Wellcome Collaborator award (214314/Z/18/Z), a JS McDonnell Foundation award (JSMF220020372), and by the Jean Francois and Marie-Laure de Clermont Tonerre Foundation. The Wellcome Centre for Integrative Neuroimaging and Wellcome Centre for Human Neuroimaging are each supported by core funding from the Wellcome Trust (203139/Z/16/Z, 203147/Z/16/Z). The Sainsbury-Wellcome centre is supported by core funding from the Wellcome Trust (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755).

References

Baldassano, C., Hasson, U., and Norman, K.A. (2018). Representation of Real-World Event Schemas during Narrative Perception. J. Neurosci. *38*, 9689–9699.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433.

Bao, X., Gjorgieva, E., Shanahan, L.K., Howard, J.D., Kahnt, T., and Gottfried, J.A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. Neuron *102*, 1066-1075.e5.

Baram, A.B., Muller, T.H., Nili, H., Garvert, M.M., and Behrens, T.E.J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. Neuron *109*, 713-723.e7.

Barron, H.C., Garvert, M.M., and Behrens, T.E.J. (2016). Repetition suppression: A means to index neural representations using BOLD? Philos. Trans. R. Soc. B Biol. Sci. *371*, 20150355.

Barry, C., Ginzberg, L.L., O'Keefe, J., and Burgess, N. (2012). Grid cell firing patterns signal environmental novelty by expansion. Proc. Natl. Acad. Sci. *109*, 17687–17692.

Bartlett, F.F.C. (1932). Remembering: An experimental and social study. Cambridge Cambridge Univ.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron *100*, 490–509.

Bongioanni, A., Folloni, D., Verhagen, L., Sallet, J., Klein-Flügge, M.C., and Rushworth, M.F.S. (2021). Activation and disruption of a neural mechanism for novel choice in monkeys. Nat. 2021 5917849 *591*, 270–274.

Bozinovski, S., and Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. Proc. Symp. Inform. *3*, 121–126.

Chadwick, M.J., Jolly, A.E.J., Amos, D.P., Hassabis, D., and Spiers, H.J. (2015). A goal direction signal in the human entorhinal/subicular region. Curr. Biol. *25*, 87–92.

Constantinescu, A.O., OReilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science (80-.). *352*, 1464–1468.

Doeller, C.F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. Nature *463*, 657–661.

Dordek, Y., Soudry, D., Meir, R., and Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. Elife *5*, 1–36.

El-Gaby, M., Harris, A.L., Whittington, J.C.R., Dorrell, W., Bhomick, A., Walton, M.E., Akam, T., and Behrens, T.E.J. (2023). A Cellular Basis for Mapping Behavioural Structure. BioRxiv 2023.11.04.565609.

Foster, D.J. (2017). Replay Comes of Age. Annu. Rev. Neurosci. *40*, 581–602.

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., and Moser, E.I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. Nature *446*, 190–194.

Garvert, M.M., Dolan, R.J., and Behrens, T.E.J. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife *6*, 1–20.

George, D., Rikhye, R. V., Gothoskar, N., Guntupalli, J.S., Dedieu, A., and Lázaro-Gredilla, M. (2021). Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. Nat. Commun. 2021 121 *12*, 1–17.

Ghosh, V.E., and Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. Neuropsychologia *53*, 104–114.

Gilboa, A., and Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated

Memory. Trends Cogn. Sci. *21*, 618–631.

Grill-Spector, K., and Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychol. (Amst). *107*, 293–321.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. Nature *436*, 801–806.

Harlow, H.F. (1949). The formation of learning sets. Psychol. Rev.

Jacobs, J., Weidemann, C.T., Miller, J.F., Solway, A., Burke, J.F., Wei, X.-X.X., Suthana, N., Sperling, M.R., Sharan, A.D., Fried, I., et al. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. Nat. Neurosci. *16*, 1188–1190.

Jezek, K., Henriksen, E.J., Treves, A., Moser, E.I., and Moser, M.-B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. Nature *478*, 246–249.

Kaefer, K., Nardin, M., Blahna, K., and Csicsvari, J. (2020). Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching. Neuron.

van Kesteren, M.T.R., Beul, S.F., Takashima, A., Henson, R.N., Ruiter, D.J., and Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. Neuropsychologia *51*, 2352–2359.

Van Kesteren, M.T.R., Rijpkema, M., Ruiter, D.J., and Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. J. Neurosci. *30*, 15888–15894.

Kriegeskorte, N., Goebel, R., and Bandettini, P.A. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. *103*, 3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

Kumaran, D., Summerfield, J.J., Hassabis, D., and Maguire, E.A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. Neuron *63*, 889–901.

Kumaran, D., Hassabis, D., and Mcclelland, J.L. (2016). What Learning Systems do Intelligent Agents Need ? Complementary Learning Systems Theory Updated. Trends Cogn. Sci. *20*, 512–534.

Liu, Y., Dolan, R.J., Kurth-Nelson, Z., and Behrens, T.E.J. (2019). Human Replay Spontaneously Reorganizes Experience. Cell.

Mack, M.L., Preston, A.R., and Love, B.C. (2020). Ventromedial prefrontal cortex compression during concept learning. Nat. Commun. 2020 111 *11*, 1–11.

Mark, S., Moran, R., Parr, T., Kennerley, S.W., and Behrens, T.E.J. (2020). Transferring structural knowledge across cognitive maps in humans and models. Nat. Commun. 2020 111 *11*, 1–12.

Mark, S., Schwartenbeck, P., Hahamy, A., Samborska, V., Baram, A.B., and Behrens, T.E. (2024). Flexible neural representations of abstract structural knowledge in the human Entorhinal Cortex. Elife *13*.

Mathis, A., Stemmler, M.B., and Herz, A.V.M. (2015). Probable nature of higher-dimensional symmetries underlying mammalian grid-cell activity patterns. Elife *2015*, 1–29.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (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. Psychol. Rev. *102*, 419–457.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Moneta, N., Garvert, M.M., Heekeren, H.R., and Schuck, N.W. (2023). Task state

representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. Nat. Commun. 2023 141 *14*, 1–21.

Morrissey, M.D., Insel, N., and Takehara-nishiuchi, K. (2017). Generalizable knowledge outweighs incidental details in prefrontal ensemble code over time. Elife *6*, 1–20.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. *10*, e1003553.

O'Reilly, R.C., and McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. Hippocampus *4*, 661–682.

Park, S.A., Miller, D.S., Nili, H., Ranganath, C., and Boorman, E.D. (2020). Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. Neuron *107*, 1226-1238.e8.

Park, S.A., Miller, D.S., and Boorman, E.D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. Nat. Neurosci. 2021 249 *24*, 1292–1301.

Piaget, J. (1948). Le language et la pensée chez l'enfant. In Delachaux et Niestlé, p.

Richards, B.A., Xia, F., Santoro, A., Husse, J., Woodin, M.A., Josselyn, S.A., and Frankland, P.W. (2014). Patterns across multiple memories are identified over time. Nat. Neurosci. *17*, 981–986.

Robertson, E.M. (2018). Memory instability as a gateway to generalization. PLOS Biol. *16*, e2004633.

Rossion, B., and 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.

Samborska, V., Butler, J.L., Walton, M.E., Behrens, T.E.J., and Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for

generalizing the structure of problems. Nat. Neurosci. 2022 2510 *25*, 1314–1326.

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. Nat. Neurosci. *16*, 486–492.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., and Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. Cell *186*, 4885-4897.e14.

Shepard, R.N. (1987). Toward a Universal Law of Generalization for Psychological Science. Science (80-.). *237*, 1317–1323.

Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat. Neurosci. *20*, 1643–1653.

Stickgold, R., and Walker, M.P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. Nat. Neurosci. 2013 162 *16*, 139–145.

Sun, W., Advani, M., Spruston, N., Saxe, A., and Fitzgerald, J.E. (2023). Organizing memories for generalization in complementary learning systems. Nat. Neurosci. 2023 268 *26*, 1438–1448.

Tacikowski, P., Kalender, G., Ciliberti, D., and Fried, I. (2024). Human hippocampal and entorhinal neurons encode the temporal structure of experience. Nat. 2024 1–8.

Takehara-Nishiuchi, K., and McNaughton, B.L. (2008). Spontaneous changes of neocortical code for associative memory during consolidation. Science (80-.). *322*, 960–963.

Takehara, K., Kawahara, S., and Kirino, Y. (2003). Time-Dependent Reorganization of the Brain Components Underlying Memory Retention in Trace Eyeblink Conditioning. J. Neurosci. *23*, 9897–9905.

Tang, W., Shin, J.D., and Jadhav, S.P. (2023). Geometric transformation of cognitive

maps for generalization across hippocampal-prefrontal circuits. Cell Rep. *42*.

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208.

Tompary, A., and Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. Neuron *96*, 228-241.e5.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. (2007). Schemas and Memory Consolidation. Science (80-.). *316*, 76–82.

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., and Morris, R.G.M. (2011). Schema-Dependent Gene Activation and Memory Encoding in Neocortex. Science (80-.). *333*, 891–895.

Veselic, S., Muller, T.H., Gutierrez, E., Behrens, T.E.J., Hunt, L.T., Butler, J.L., and Kennerley, S.W. (2023). A cognitive map for value-guided choice in ventromedial prefrontal cortex. BioRxiv 2023.12.15.571895.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage *137*, 188–200.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., and Botvinick, M.M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. *21*, 860–868.

Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. Cell *183*, 1249-1263.e23.

Whittington, J.C.R., McCaffary, D., Bakermans, J.J.W., and Behrens, T.E.J. (2022). How to build a cognitive map. Nat. Neurosci. 2022 2510 *25*, 1257–1272.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., and Nichols, T.E. (2014). Permutation inference for the general linear model. Neuroimage *92*, 381–397.

Yoon, K., Buice, M.A., Barry, C., Hayman, R., Burgess, N., and Fiete, I.R. (2013). Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nat. Neurosci. *16*, 1077–1084.

Yu, J.Y., Liu, D.F., Loback, A., Grossrubatscher, I., and Frank, L.M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory. Nat. Commun. 2018 91 *9*, 1–11.

Zeithamova, D., and Preston, A.R. (2010). Flexible Memories: Differential Roles for Medial Temporal Lobe and Prefrontal Cortex in Cross-Episode Binding. J. Neurosci. *30*, 14676–14684.