Map making: Constructing, combining, and navigating abstract cognitive maps

Seongmin A. Park[1,2], Douglas S. Miller[1,2], Hamed Nili[3], Charan Ranganath[2,4], Erie D. Boorman[1,4]

1. Center for Mind and Brain, University of California, Davis, USA
2. Center for Neuroscience, University of California, Davis, USA
3. Wellcome Centre for Integrative Neuroimaging, University of Oxford, UK
4. Department of Psychology, University of California, Davis, USA

Lead contact: S. A. Park (apark@ucdavis.edu) and E. D. Boorman (edboorman@ucdavis.edu)

ABSTRACT

Cognitive maps are thought to enable model-based inferences from limited experience that can guide novel decisions–a hallmark of goal-directed behavior. We tested whether the hippocampus (HC), entorhinal cortex (EC), and ventromedial prefrontal cortex (vmPFC)/medial orbitofrontal cortex (mOFC) organize abstract and discrete relational information into a cognitive map to guide novel choices. Subjects learned the status of people in two separate unseen 2-D social hierarchies defined by competence and popularity piecemeal from binary comparisons. Although only one dimension was ever behaviorally relevant, multivariate activity patterns in HC, EC and mOFC were linearly related to the Euclidian distance between people in the mentally reconstructed 2-D space. Hubs created unique comparisons between the two hierarchies, enabling inferences between novel pairs of people. We found that both behavior and neural activity in EC and vmPFC/mOFC reflected the Euclidian distance to the retrieved hub, which was reinstated in HC. These findings reveal how abstract and discrete relational structures are represented, combined, and navigated in the human brain.

INTRODUCTION

To form rich world models, sparse observations often sampled from separate experiences need to be integrated into a coherent representation. There has been a recent surge of interest in the long-standing theory that the hippocampus (HC) and entorhinal cortex (EC) may organize spatial and non-spatial relational information into such a 'cognitive map' for goal-directed behavior (Behrens et al., 2018; Bellmund, Gärdenfors, Moser, & Doeller, 2018; Constantinescu, O'Reilly, & Behrens, 2016; Howard Eichenbaum & Cohen, 2014; Ekstrom & Ranganath, 2018; Hafting, Fyhn, Molden, Moser, & Moser, 2005; Moser, Kropff, & Moser, 2008; O'Keefe & Nadel, 1978; Schiller et al., 2015; Schuck, Cai, Wilson, & Niv, 2016; Tolman, 1948; Wikenheiser & Schoenbaum, 2016). While past studies have identified neural signals in the HC and EC indicative of a cognitive map primarily using continuous task dimensions with online sensory feedback during task performance (e.g. visual, auditory, vestibular) (Aronov, Nevers, & Tank, 2017; Bao et al., 2019; Constantinescu et al., 2016; Doeller, Barry, & Burgess, 2010; Howard Eichenbaum & Cohen, 2014; Hafting et al., 2005; Nau, Navarro Schröder, Bellmund, & Doeller, 2018; O'Keefe & Nadel, 1978; Theves, Fernandez, & Doeller, 2019), many important everyday decisions involve discrete entities that vary along multiple abstract dimensions that are sampled piecemeal, one experience at a time, in the absence of continuous sensory feedback, such as with whom to collaborate or where to eat. How the brain constructs a cognitive map of abstract relationships between discrete entities from piecemeal experiences is unclear.

A powerful advantage of a cognitive map of an environment or task is the ability to make inferences from sparse observations that can dramatically accelerate learning and even guide novel decisions never faced before (Banino et al., 2018; Behrens et al., 2018; Jones et al., 2012; Stachenfeld, Botvinick, & Gershman, 2017; Tolman, 1948; Vikbladh et al., 2019), a hallmark of behavioral flexibility and a key challenge in artificial intelligence (Behrens et al., 2018; Kriete, Noelle, Cohen, & O'Reilly, 2013; Wang et al., 2018). This is in part because a cognitive map of a task space allows "shortcuts" and "novel routes" to be inferred, as in physical space. To provide a concrete example, understanding the structure of family trees allows one to infer new relationships, such as the following: because Sally is John's sister and Sue is John's daughter, Sue must be Sally's niece without ever directly learning this relationship. Biologically inspired computational models show the map-like coding schemes found in the HC and EC can in principle enable agents to perform vector navigation, including planning new routes and finding shortcuts to a goal in physical space (Banino et al., 2018; Bush, Barry, Manson, & Burgess, 2015; Whittington, Muller, Mark, Barry, & Behrens, 2018). In particular, so-called place cells in HC and

grid cells in medial EC implement neural codes that permit calculation of predicted position (Moser et al., 2008; O'Keefe & Nadel, 1978; Stachenfeld et al., 2017), direction (Banino et al., 2018; Chadwick, Hassabis, Weiskopf, & Maguire, 2010), and Euclidian distance (Behrens et al., 2018; Bellmund et al., 2018; Howard et al., 2014) in physical space. Here, we asked whether similar neural computations might underlie novel *non-spatial* inferences about discrete entities organized in an abstract 2-D space that could only be re-constructed from piecemeal experience from binary comparisons on one dimension at a time, learned on separate days.

A parallel literature based on recent studies focusing on the orbitofrontal cortex (OFC) has motivated a related theory that the OFC represents one's current position in a cognitive map, not of physical space, but of task space (Schuck et al., 2016; Takahashi et al., 2017; Walton, Behrens, Buckley, Rudebeck, & Rushworth, 2010; Wikenheiser & Schoenbaum, 2016; Wilson, Takahashi, Schoenbaum, & Niv, 2014). Recent findings further suggest a specialized role for mOFC in representing all the latent (or perceptually unsignaled) components of the task space that define one's current state in the task (Schuck et al., 2016; Wilson et al., 2014). This function is proposed to play an important role both during learning (Takahashi et al., 2017; Walton et al., 2010) and choice (Jones et al., 2012; Stalnaker, Cooch, & Schoenbaum, 2015). Recent studies have indeed discovered that the OFC represents latent task states during learning and choice in support of this theory (Chan, Niv, & Norman, 2016; Schuck et al., 2016; Wikenheiser, Marrero-Garcia, & Schoenbaum, 2017), yet to our knowledge there has been little direct evidence of map-like representations (e.g. position, direction, or distance) of the task space in OFC. Moreover, whether this proposed OFC function would extend to representing a cognitive map of an abstract social space, or whether it would instead transfer to areas implicated in social cognition is unclear.

In addition to *representing* cognitive maps, both the HC and OFC have been implicated in model-based *inference*, such that distinct items, or stimuli and rewards, that were not directly associated can be associated or integrated through an overlapping, shared associate (Jones et al., 2012; Koster et al., 2018; Kurth-Nelson, Economides, Dolan, & Dayan, 2016; Schlichting & Preston, 2014; Tompary & Davachi, 2017; Wimmer & Shohamy, 2012). Computational models have proposed how a unitary mechanism in the HC system could additionally underlie transitive inferences about ordinal rank (Koster et al., 2018; Kumaran & McClelland, 2012) (though see (Frank, Rudy, Levy, & O'Reilly, 2005)). In addition to demonstrations that HC is necessary for transitive inferences (Howard Eichenbaum, Otto, & Cohen, 1996), studies in animal models have demonstrated that the OFC is necessary for model-based inferences, but not for decisions based on directly learned cached values (Jones et al., 2012). While these HC and OFC roles for associating or integrating individual items have been documented, how these proposed functions

relate to the construction of broader cognitive maps beyond elemental associations has been elusive. In particular, it is possible similar mechanisms enable not only the chaining of associations, but also the integration of distinct relational structures into a larger cognitive map from sparse observations (e.g. the integration of family trees through marriage) that even respects metric relationships (e.g. vector directions and distances).

We tested two alternative hypotheses concerning how the human brain could represent and flexibly switch between different behaviorally relevant dimensions that characterize the same entities to guide inferences. First, if a neural representation depends on the current behaviorally relevant dimension alone, then we would predict preferential encoding of that dimension in brain areas important for current behavior. On the other hand, if relationships between people are projected into a unitary space defined by their respective values on two independent dimensions, then we would predict a single neural representation such that behavioral and neural activity reflect the distance over a 2-D Euclidian space between entities, rather than the behaviorally relevant 1-D rank alone.

RESULTS

Participants learned relational maps of two 2-D social hierarchies and used hubs between them to make inferences between novel pairs of individuals

To investigate whether the human brain constructs a cognitive map of multidimensional social hierarchies and leverages this map for model-based inference during decisions, we asked participants to learn the status of unfamiliar people in two separate groups organized hierarchically on two orthogonal dimensions: competence and popularity (**Fig. 1A**). Importantly, participants never saw the 1- or 2-D hierarchies. Instead, they were able to learn the relative ranks of neighboring people who differed by only one level on one dimension at a time through a series of feedback-based dyadic comparisons (see (Kumaran, Melo, & Duzel, 2012)). Participants could then infer the relative ranks of non-neighbors in their group through transitive inference. To learn the two hierarchies, participants completed three days of behavioral training, with a 48-hour gap between each training session, and on the last day, the fMRI experiment (**Fig. 1C**).

During the first two days of training, participants learned the relative status of two groups of 8 “entrepreneurs” separately, on only one dimension per day (**Supplementary Fig.1A** for day 1 and **Supplementary Fig.1B** for day 2 training), such that people in a group were only compared against others belonging to the same group (two groups of 8 entrepreneurs, **Fig.1A**). At the end of day 2, test trials without feedback ensured subjects could make transitive inferences to determine the status of remaining members within a group, on each dimension separately (test2 in **Supplementary Fig.1B**), indicating they had learned the two 1-D hierarchies for both groups. Importantly, participants were never asked to combine the two dimensions in either group. We included four rank levels per dimension to ensure that differences between rank levels 2 and 3 could not simply be explained by differences in win frequency, since these people each “won” and “lost” on ½ of trials. A separate behavioral experiment consisting of a placement task and a ratings task after day 3 training conducted on a separate group of participants showed that they had successfully learned the four levels for each dimension in the social hierarchy, and importantly, could accurately differentiate between rank levels 2 and 3 for both dimensions (**Supplementary Fig.5**). For the third day of training, fMRI participants learned from select between-group comparisons for the first time (**Supplementary Fig.2A**). That is, participants only learned the relative rank of selected entrepreneurs in each group referred to as ‘hubs’, who were paired against both group members (**Supplementary Fig.2B**). By limiting between-group comparisons only to hubs, we were able to create comparative paths connecting each of the individuals in

different groups, which could be leveraged to perform model-based inferences between novel pairs of entrepreneurs between groups.

We analyzed fMRI data acquired from twenty-seven subjects who successfully learned the relative ranks of entrepreneurs in each of the two social hierarchies (> 85% performance criterion for inferring relative status of each group member in both dimensions, tested on Day 2 training). In each trial of an fMRI block, participants were asked to make a binary decision about who was higher rank in one *or* the other dimension between the first face (F1) and the second face (F2) presented sequentially (**Fig. 1B**). Unbeknownst to participants, individuals who were presented at the time of F1 and F2 were selected from different groups and were not hubs in the given dimension (non-hubs; **Supplementary Fig.2D**), meaning they had not been previously compared. Following decisions, a third face (F3) was presented, and participants were asked to perform a cover task to simply indicate their gender. F3 was presented from among hubs (**Supplementary Fig.2E**) to test for hypothesized fMRI suppression of the relevant latent hub, relative to other matched but non-relevant hubs, that may have been retrieved from memory to guide model-based inferences.

Participants never learned the relative rank of F1 and F2. Instead, they could make a model-based inference through one of two specific hub individuals who had previously been paired with both F1 and F2 during training. Successful inferences, therefore, could rely on building an internal representation of the social hierarchies and a relational memory of the relative positions of F1, F2, and the hub. We predicted that the inference is made along a trajectory in abstract space connecting the two individuals via the (unseen) hub. The hubs were two individuals (H1 and H2) who had been paired with both individuals (F1 and F2) in a task-relevant dimension. The H1 (H2) is uniquely paired with F1 (F2) in between-group comparison while it belongs to the same group with F2 (F1). These task-relevant hubs in one dimension differ from those in the other dimension, which means that to make an accurate inference of the relative status of the same pair of individuals in the two different dimensions, participants needed to retrieve different hubs, which would alter the inference trajectories (**Supplementary Fig.2F**). Since the inference trajectory is anchored by the position of the hub, we were able to track the putative trajectory used by participants by examining which hub between H1 and H2 was selectively retrieved during inferences. Furthermore, we could examine whether participants utilize only the task-relevant rank distance (D) or also the Euclidean distance (E) between individuals' positions in the cognitive map.

Participants were able to successfully infer the relative position of novel pairs of individuals (overall accuracy ± standard error mean (s.e.m.) = 93.6±0.77%). Nonetheless, the shorter the distance between individuals, the more difficult the decision about relative positions in the

hierarchy. To examine the effects of distance of potential trajectories on decision making, we regressed choice reaction times (RT) on different distance measures using a multiple linear regression model, thereby allowing them to compete to explain RT variance. We included the Euclidean distance from the hub (H2) to F1 (E_{H2F1}), the Euclidean distance from the other hub (H1) to F2 (E_{H1F2}), the relative rank in the task-relevant dimension, which is the 1-D distance between H2 and F1 (D_{H2F1}), the 1-D distance between H1 and F2 (D_{H1F2}), (**Fig. 1D**), as well as both 1-D and 2-D distances between F1 and F2 (D_{F1F2} and E_{F1F2}, respectively), (**Fig. 1E**) to control for their possible covariation with hub-related distances. We found that the greater the 1-D and 2-D Euclidian distance between F1 and H2 (D_{H2F1} and E_{H2F1}, respectively), the faster the RT (βD_{H2F1}±sem=-52.9±11.9, t_{26}=-4.5, p=4.5e-05; βE_{H2F1}±sem=-49.4±5.7, t_{26}=-8.8; p=0.003), in addition to an effect of the 1-D distance between F1 and F2 (D_{F1F2}) (βD_{F1F2}±sem=-64.6±12.5, t_{26}=-5.3, p=0.0002), (**Fig. 1F**). Moreover, we found that E_{H2F1} accounted for variation in RTs better (t_{26}=-2.73, p=0.011, paired t-test) than E_{H1F2} (βE_{H1F2}=13.5±5.5, t_{26}=1.0, p=0.33). Our behavioral results show that participants preferentially recall H2 as the task-relevant hub to aid in the comparison between novel pairs of faces, with the Euclidian distance to H2 explaining variance over and above the 1-D distance alone.

Neural activity reflects the Euclidian distance to the retrieved hub during inferences

To examine whether neural activity during choices was likewise modulated by the distance of inference trajectories via the hub over the Euclidean space, we regressed whole brain activity at the time of the inference (F2) against a general linear model (GLM1) that included the parametric regressors E_{H2F1}, E_{H1F2}, and the (cosine) vector angles between the hub and the face (A_{H2F1} and A_{H1F2}), (see Methods for GLM1 specification). We found neural correlates of E_{H2F1} in *a priori* predicted areas of bilateral EC (p_{SVC}<0.01 at the peak level, [x,y,z]=[20,2,-26], t_{26}=4.17, p_{SVC}=0.031 for right EC; [x,y,z]=[-18,-10,-34], t_{26}=4.43, p_{SVC}=0.017 for left EC, small volume correction (SVC) was applied at the peak level in anatomically defined ROIs (Amunts et al., 2005; Zilles & Amunts, 2010)), ventral medial prefrontal cortex encompassing medial OFC (vmPFC/mOFC) (peak voxel [x,y,z]=[2,32,-6], t_{26}=4.82, p_{TFCE}<0.05), and in the right lateral OFC (lOFC, [x,y,z]=[30,34,-18], t_{26}=4.12, p_{TFCE}<0.05), (**Fig. 2A**). Note that we apply whole-brain threshold-free cluster enhancement (TFCE) correction at p_{TFCE}<0.05 for all analyses, unless specifically stated for our *a priori* hypotheses where we may apply small-volume correction (SVC) at the peak level. For a full list of brain areas surviving outside of these predicted areas that survive TFCE correction at p_{TFCE}<0.05, see **Supplementary Table1A**. No significant effects were

found for the alternative terms E_{H1F2}, A_{H2F1} and A_{H1F2} at these thresholds (**Supplementary Fig.7**). These analyses show that the EC, vmPFC/mOFC, and lOFC compute or use the Euclidian distance to the context-relevant latent hub H2 to guide inference decisions based on a cognitive map.

We also tested our competing hypothesis that the brain flexibly switches between behaviorally relevant and irrelevant dimensions with simultaneous coding of both dimensions, but in different brain regions. Specifically, we tested whether the current behaviorally relevant rank distance (D_{H2F1}) and the behaviorally irrelevant rank distance (I_{H2F1}) better explain neural activity in the EC and vmPFC/mOFC, or elsewhere in the brain (GLM2), than the Euclidian distance (E_{H2F1}). It is important to note that E_{H2F1} can be decomposed into D_{H2F1} and I_{H2F1} (with equal weighting), so if E_{H2F1} in fact is reflected by neural activity in a region, we would expect effects of both D_{H2F1} and I_{H2F1} in the same region.

We found a largely overlapping network of areas to those reported for E_{H2F1} above showing positive effects of both D_{H2F1} and I_{H2F1} at a reduced threshold (**Fig. 2A** and **Supplementary Table1B**), including vmPFC/mOFC (D_{H2F1}: [x,y,z]=[-2,32,-6], t_{26}=5.25, p_{SVC}=0.003, and [x,y,z]=[4,30,-6], t_{26}=4.96, p_{SVC}=0.005; I_{H2F1}: [x,y,z]=[6,36,-6], t_{26}=3.73, p_{SVC}=0.060, and [x,y,z]=[-4,30,-4], t_{26}=3.48, p_{SVC}=0.097 corrected at the peak level in *a priori* anatomically defined ROIs in vmPFC/mOFC (Neubert, Mars, Sallet, & Rushworth, 2015)), and the EC (D_{H2F1}: [x,y,z]=[18,-12,-26], t_{26}=5.01, p_{SVC}=0.021; I_{H2F1}: [x,y,z]=[30,-14,-32], t_{26}=3.14, p_{SVC}=0.090, in *a priori* EC ROIs (Amunts et al., 2005; Zilles & Amunts, 2010)). A statistical conjunction analysis (Friston, Penny, & Glaser, 2005) between D_{H2F1} and I_{H2F1} showed common effects in the vmPFC/mOFC ([x,y,z]=[6,36,-6], t_{26}=3.74, p_{SVC}=0.033, and [x,y,z]=[-2,32,-6], t_{26}=3.54, p_{SVC}=0.054) and the right EC ([x,y,z]=[20,-18,-26], t_{26}=2.90, p_{SVC}=0.067), (**Fig. 2C**). We did not find any significant effects of D over I and I over D in the vmPFC/mOFC and EC. Importantly, we did not find evidence D was encoded in one set of brain regions and I was simultaneously encoded in a different set of brain regions (**Supplementary Fig.4**). We likewise did not find any significant effects of 1-D distances from the alternative hub, H1 (D_{H1F2} and I_{H1F2}; **Supplementary Fig.7**). These findings support the interpretation that vmPFC/mOFC and EC encode E_{H2F1}, which is composed of both D and I (with equal weighting).

To formally arbitrate between different possible decision trajectories, we used Bayesian model selection (BMS) to compare 2-D and 1-D metrics for different possible trajectories (or comparisons) through the hub H1 (E_{H1F2}, D_{H1F2}, and I_{H1F2}), those through the hub H2 (E_{H2F1}, D_{H2F1}, and I_{H2F1}), and also direct distances between F1 and F2, rather than trajectories via the hub (E_{F1F2}, D_{F1F2}, and I_{F1F2}). In addition, different hypothetical cognitive spaces may have different underlying

metrics. That is, if participants do not construct a cognitive map in a 2-D Euclidean space, but rather a different architecture for representing social hierarchies, alternative distance metrics may better account for the neural data than the Euclidean distance. For example, if inferences were made through the sequential retrieval of individuals linking F1 to F2, activity in EC and vmPFC/mOFC should be better explained by the shortest number of links (L; note this is equivalent to $1+D_{H2F1}$). Alternatively, if a cognitive map encodes only the vector angle between individuals in a polar coordinate system, neural activity should encode the angle between individuals (A_{F1F2}, A_{F1H1} and A_{F2H1}), while it should be invariant to the length (**Supplementary Fig.6A**). This formal comparison revealed clear evidence in favor of the Euclidian distance through hub H2 (E_{H2F1}) in the EC and vmPFC/mOFC, supporting the hypothesis that the relevant latent hub H2 is used for model-based inference using a cognitive map in Euclidian space (exceedance probability=0.82 in left EC; 0.91 in right EC; 0.89 in left vmPFC/mOFC;0.85 in right vmPFC/mOFC; **Fig. 2B**; **Supplementary Table2**). Taken together, our findings show that EC and vmPFC/mOFC compute or utilize Euclidian distances over the 2-D social space to guide decisions.

HC reinstates the hub to guide inferences

The analyses presented so far imply the context-relevant hub is retrieved from memory to guide inferences. We therefore searched for neural evidence of a reinstatement of the latent hub along this trajectory to guide decisions. To address this question, we adopted a variant of repetition suppression (RS), but for a retrieved rather than explicitly presented item. Notably, RS has been proposed as a means to assess the information content of neuronal ensembles in the human brain (Barron, Garvert, & Behrens, 2016; Boorman, Rajendran, O'Reilly, & Behrens, 2016; Grill-Spector, Henson, & Martin, 2006; Klein-Flugge, Barron, Brodersen, Dolan, & Behrens, 2013). During F3 presentation, participants were exposed to one of eight hub individuals (**Supplementary Fig.2E**). We hypothesized that if the relevant hub that bridges F1 and F2 in the given dimension is presented during F3 presentation, directly after participants retrieve the relevant hub, then the BOLD signal in areas reinstating that hub should be suppressed compared to other trials presenting matched but non-relevant hubs. We included only hubs as F3 because these individuals are equally matched for win/loss frequency (each winning on ½ of trials and losing on the other ½) and experience (i.e. presentation frequency), thereby ruling out these potential confounding factors. Moreover, we ensured the Euclidean distance from F2 to F3 (E_{F2F3}) was not different when F3 was H1, H2, or a non-relevant hub (F_2=0.77, p=0.47, one-way ANOVA;

Supplementary Fig.3C), in order to control for the distance between presented faces for each type of hub.

We found that the right HC (peak voxel [x,y,z]=[38,-22,-12], t_{26}=3.41, p_{SVC}=0.021 corrected in a small volume based on an anatomically defined HC ROI (Yushkevich et al., 2015)) was the only brain area showing greater suppression, specifically for the relevant H2 presentations (β=-0.46±0.13) compared to all non-relevant hub presentations (β=-0.19±0.11; t_{26}=4.54, p<0.01, paired *t*-test in the independent, anatomically defined ROI), (**Fig. 3**). We did not find any brain area showing greater suppression during presentation of the other possible hub, H1 (β=-0.02±0.21; t_{26}=0.81, p=0.43), consistent with our analyses reported above, indicating that participants wait for the presentation of F2 to make a backward inference about its rank relative to F1 by preferentially retrieving H2. Importantly, we did not find evidence for decreasing (or increasing) activity in the HC when we modeled whole brain activity as a function of Euclidean distance from the hub to F3 (E_{H2F1} and E_{H1F2}) at the time of F3 presentation (**Supplementary Fig.3D**), further suggesting that HC suppression was specific to the latent hub itself, rather than driven by proximity in the Euclidean space, thus ruling out a distance-based suppression account between presented faces (Garvert, Dolan, & Behrens, 2017). Taken together, these findings show that HC reinstates the behaviorally relevant hub to guide model-based inferences between distinct relational structures.

HC, EC, and vmPFC/mOFC represent social hierarchies in a 2-D space

To directly examine the cognitive map's representational architecture, we measured the pattern similarity between different face presentations during F1 and F2. Under the hypothesis that more proximal positions in the cognitive map will be represented by increasingly similar patterns of neuronal activity, we used representational similarity analysis (RSA) to test the extent to which patterns of activity across voxels in the HC, EC and vmPFC/mOFC are linearly related to the Euclidian distance between faces in the true 4-by-4 social network. We reasoned that if the cognitive map of social networks is characterized by two independent dimensions projected into a Euclidean space, the level of dissimilarity between neural representations evoked by each face (**Fig. 4A**) should be explained by pairwise Euclidean distances (E), in addition to the pairwise rank differences in the task-relevant dimension (D) (**Fig. 4B**).

In hypothesis-driven analyses, we first analyzed data from *a priori* selected anatomical ROIs (**Fig. 4C**), including the bilateral HC (Yushkevich et al., 2015), EC (Amunts et al., 2005; Zilles & Amunts, 2010), and vmPFC/mOFC (Neubert et al., 2015). The representational

dissimilarities estimated in the ROIs was explained both by the model representational dissimilarity matrix (RDM) of pairwise Euclidian distance in 2-D space (E, one-sided Wilcox signed rank test, df=26, p_{FWE}<0.01, Holm-Bonferroni correction for multiple comparisons across numbers of model RDMs and ROIs) and, in a separate RDM, by the pairwise rank difference in the task-relevant distance (D, p_{FWE}<0.01) between individuals (**Fig. 4D**). Based on previous demonstrations that univariate amygdala activity (Kumaran et al., 2012) and gray matter density (Bickart, Wright, Dautoff, Dickerson, & Barrett, 2011; Noonan et al., 2014; Sallet et al., 2011), correlate with social dominance status, we also tested anatomically defined amygdala ROIs (Tzourio-Mazoyer et al., 2002). The amygdala pattern similarity was neither explained by E nor by D, even at a reduced threshold (p>0.05, uncorrected). As a control region, we also tested the pattern similarity in primary motor cortex (Glasser et al., 2016), which was not explained by either predictor (p>0.05, uncorrected), (**Fig. 4D)**. On the other hand, the pattern similarity in HC, EC, and vmPFC/mOFC was not explained by the behavioral "context" of the task-relevant dimension (C, defined as popularity or competence trials), nor whether individuals belonged to the same group or not during training (G), (p_{FWE}>0.05), (**Fig.4D** and **Supplementary Table3A**). Importantly, the predictor of pairwise Euclidean distance (E) still significantly accounted for the pattern similarity in HC, EC, and vmPFC/mOFC (Rank correlation τ_A=0.045±0.005 for HC; τ_A=0.027±0.007 for EC; τ_A=0.048±0.006 for vmPFC/mOFC; p_{FWE}<0.01; **Supplementary Fig.8B**) after partialing out its partial correlation with rank distance (D) to ensure that D alone was not driving the pattern similarity effects (**Supplementary Fig.8A**). To confirm that the pattern similarity truly reflected E, we tested for separate effects of D and I. Decomposing E into the terms D (**Supplementary Fig.8C**) and I (**Supplementary Fig.8D**) revealed a linear relationship between pattern similarity and both distance components. These analyses show that, in addition to D, I contributed significantly to representations in these regions, supporting the interpretation that the social hierarchy was represented in 2-D, even though only one dimension was behaviorally relevant.

Notably, the effect of E was strongest for within-group pairs (i.e. individuals who were part of the same group during training; **Supplementary Fig.9C**) and for between-group pairs involving hubs (i.e. individuals and their hubs who were compared during between-group learning in day 3 training; **Supplementary Fig.9D**) and weaker for never-compared between-group pairs of non-hubs (**Supplementary Fig.9E**) in the bilateral HC, EC and vmPFC/mOFC (p_{FWE}<0.01, two-sided Wilcoxon signed-rank test; **Supplementary Fig.9F** and **Supplementary Table3C**; the mean rank correlations are shown in **Supplementary Fig.9F** and **Supplementary Table3B**). Notably, however, the effect of E was still significant, though weaker, for never-compared between-group

pairs of non-hubs in HC alone, suggesting that HC integration may lead EC and vmPFC/mOFC. This pattern is consistent with the interpretation that the 2-D representations of the two different groups had not yet been fully combined into a single combined cognitive map, but rather two partly separate group-dependent maps, and that subjects therefore had to utilize the hubs for between-group inferences.

In addition to these hypothesis-driven analyses, we also performed whole-brain exploratory analyses to test whether the neural representation of the social network extends to a broader set of regions. Specifically, we measured the extent to which each predictor (the model RDM of E and D) explains the pattern similarity measured from searchlight-based pattern analyses across the whole brain. This analysis revealed that the pairwise Euclidean distance (E) significantly explained the representational similarity between faces in HC and EC, as shown by the ROI analyses, and also in medial, central, and lateral OFC, among other areas ($p_{TFCE}<0.05$; **Fig. 4F** and **Supplementary Table4A**). A separate RDM based on the pairwise 1-D rank distance (D) significantly explained representational similarity between activity patterns in the lateral OFC, medial prefrontal cortex (mPFC), and posterior cingulate cortex (PCC) ($p_{TFCE}<0.05$; **Fig. 4G** and **Supplementary Table4B**). Furthermore, partialing out D from the RDM for E revealed significant effects in these same areas of HC, EC, and central/medial OFC, confirming that these representations were not simply driven by D alone (**Fig. 4H and Supplementary Table4C**). Our findings suggest that the HC-EC and vmPFC/mOFC do not treat dimensions separately when representing individuals in a social network space. Instead, representations vary along the multidimensional cognitive map even when only one dimension is relevant to current behavioral goals.

DISCUSSION

The HC formation is thought to contain relational codes of our experiences (Howard Eichenbaum & Cohen, 2014). Memories of place and their spatial relationship are key elements to constructing a cognitive map of physical space (Butler, Hardcastle, & Giocomo, 2019; Kropff, Carmichael, Moser, & Moser, 2015; Moser et al., 2008). In humans, the ability to construct an accurate cognitive map of relationships between abstract and discrete information is proposed to be critical for high-level model-based decision making and generalization (Behrens et al., 2018; Bellmund et al., 2018; Vikbladh et al., 2019). We show that the HC and EC, which are famously known for their key roles in the ability to navigate physical space (Moser et al., 2008; O'Keefe & Nadel, 1978) and simultaneously their roles in episodic memory (H. Eichenbaum, Yonelinas, & Ranganath, 2007; Ekstrom & Ranganath, 2018), contribute in a more general way to the organization and 'navigation' of social knowledge in humans. Although participants were never asked to combine the two social dimensions, we found that the brain spontaneously represents individuals' status in social hierarchies in a map-like manner in 2-D space. Such a cognitive map can be used to compute routes through the 2-D space and corresponding distances (Behrens et al., 2018), which we found were computed or used to guide inferences in EC and interconnected vmPFC/mOFC, a region known to be important for value-based decision making (Boorman, Behrens, Woolrich, & Rushworth, 2009; FitzGerald, Seymour, & Dolan, 2009; Hunt et al., 2012; Lim, O'Doherty, & Rangel, 2011; Nicolle et al., 2012; Papageorgiou et al., 2017; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011; Strait, Blanchard, & Hayden, 2014). Moreover, our results show that the HC-EC system did not selectively represent the task-relevant information in our task, but the relative positions in the multidimensional space. More broadly, these findings support the HC-EC system's role in representing a cognitive map of abstract and discrete spaces to guide novel decisions that relied on that cognitive map.

We found that during model-based inferences, RTs and neural activity in EC, vmPFC/mOFC, and lOFC reflected the Euclidian distance of navigational trajectories on 2-D space via relevant hubs, over and above the behaviorally-relevant 1-D ranks. Notably, cells encoding the distance to "goals" have been documented in EC and mPFC of rodents (Boccara, Nardin, Stella, O'Neill, & Csicsvari, 2019; Butler et al., 2019; Guise & Shapiro, 2017) when navigating in physical space, while the rodent OFC has been shown to be necessary for model-based inferences (Jones et al., 2012). A separate body of work in humans and monkeys during value-based choice has consistently found value comparison signals thought to guide goal-based choices in vmPFC/mOFC (Boorman et al., 2009; FitzGerald et al., 2009; Hunt et al., 2012; Lim et

al., 2011; Nicolle et al., 2012; Papageorgiou et al., 2017; Rushworth et al., 2011; Strait et al., 2014). Taken together, this suggests the EC, vmPFC/mOFC, and lOFC compute or utilize distance computations derived from a cognitive map to guide model-based decisions.

These trajectories imply that people retrieve the task-relevant hub (H2) to guide backward inferences during novel comparisons between people from different groups. Consistent with this interpretation, suppression analyses revealed that the HC was the only brain region showing a reinstatement of this same unseen hub (H2) at the time of inference decisions, in order to link two individuals whose positions were learned in different social groups. This effect was specific to H2 (relative to both non-relevant but matched hubs and H1). Importantly, this H2-specific suppression in the HC could not be explained by the Euclidean distances between F2 and F3 (**Supplementary Fig.3C**) or H2 and F3 (**Supplementary Fig.3D**), nor by win/loss frequency, nor the degree of experience with each face, since all faces presented as F3 were hubs that were carefully matched for presentation frequency/familiarity and win/loss history. Note that making inferences via H2 does not indicate that participants integrated members in one specific group to the pre-existing structure of the other group. Considering that participants were also presented all F1-F2 pairs in reverse order, our results indicate that a different hub was preferentially recalled during inferences about the same pairs of individuals when they were presented in reverse order. Given the fact that participants were never informed that F1 and F2 were selected from non-hubs in different groups, it appears to be a more natural decision to select the hub to guide inferences after knowing both F1 and F2.

We found that the pattern similarity between faces in HC, EC, and in vmPFC/OFC was robustly and linearly related to the true Euclidian distance between faces in the 4-by-4 social network, such that closer faces in the abstract space were represented increasingly more similarly. This finding is striking for two reasons: first, the two dimensions never had to be combined to perform the task accurately; and second, the true structure was never shown to participants, but had to be reconstructed piecemeal from the outcomes of binary comparisons between neighbors in each dimension separately learned on separate days. There are several strategies that could, in principle, be used to solve this task that do not rely on a 2-D representational space. For example, the task-relevant and irrelevant dimensions could be represented in separate brain areas, a hypothesis for which we did not find clear support. Alternatively, one could envision each person's rank is represented by a linear (or logarithmic) number line, such as that found in the bilateral intraparietal area (Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004). That the neural representation in the HC, EC, and vmPFC/OFC areas automatically constructed the 2-D relational structure in our tasks instead suggests that the brain may project people, or perhaps any entities,

into a multi-dimensional cognitive or relational space such that the entity's position is defined by the feature values on each dimension (Bellmund et al., 2018). It is unclear from our study whether this finding is specialized to representing people, who are likely to be ecologically perceived as coherent entities over time, and characterized by multiple attributes, or more general to representing any entity. Precisely how this construction takes place will be an exciting topic for future studies.

Recent studies suggest that even relationships between social entities may be organized as a cognitive map in HC. In particular, these studies suggest that the HC not only represents the position of others in physical space through the firing fields of "social place cells" (Danjo, Toyoizumi, & Fujisawa, 2018; Omer, Maimon, Las, & Ulanovsky, 2018), but also their 1-D rank in a dominance hierarchy, with greater HC BOLD activity for higher rank people (Kumaran et al., 2012), self-other differences in preference-based ratings (Kaplan & Friston, 2019), and their egocentric relative direction in a space characterized by power and affiliation dimensions (Tavares et al., 2015). This latter study showed that univariate HC BOLD activity is modulated by the angle of the vector to the relative position of another person with whom subjects interacted during a role-playing game (Tavares et al., 2015). Notably, the cosine angle, unlike the Euclidean distance, is invariant to the distance to others. While this important study first suggested the HC encodes the egocentric direction of another's relative position during online social interactions, the nature of HC's representational architecture, and any putative role of the EC and vmPFC/mOFC in representing a cognitive map of social space and guiding novel inferences were previously unaddressed. In addition to dominance, theories in social cognition propose orthogonal psychological dimensions of competence and warmth along which humans represent other people in an abstract space (Fiske, 1992). We found that 2-D social hierarchies defined by independent social dimensions were represented and navigated as a cognitive map to guide novel inferences. Our study further elucidates how even social relationships are represented as a cognitive map and how that map guides novel inferences about social status.

Another recent study (Tang et al., 2019) in which human participants learned the association between novel visual stimuli and reward values over multiple days used a linear support vector machine to show that the brains of fast learners were more likely to use an efficient coding scheme to represent stimuli. Their study suggested that multidimensional coding in the brain helps participants to discriminate different stimuli while simultaneously embedding the stimuli in an efficient low-dimensional task-relevant structure. In the current task in which the rank positions of individuals in both dimensions were used to make inferences about the social hierarchical status of novel pairs of individuals, our results suggest that the accurate

representation of multidimensional task-relevant information may be critical for successful model-based decision-making.

While we found strong effects of the Euclidian distance between patterns of activity evoked by pairs of people across the entire social network, *post-hoc* comparisons suggested that these effects were strongest for pairs of faces within the same training group, and for pairs of faces and their hubs, while present (at least in HC) but weaker for pairs of faces between groups (**Supplementary Fig.9**). This pattern is consistent with the interpretation that people had formed two 2-D cognitive maps, one per hierarchy or group, and were in the process of combining these into one map of the integrated social hierarchy. This pattern also dovetails with our finding that people utilized hubs to guide inferences between pairs in different groups, rather than relying solely on the vector distances between faces in different groups. Notably, this observation is consistent with the view that the training episode may have constituted a context during which each relational map was initially formed separately for each group in HC, EC, and vmPFC/mOFC (Diana, Yonelinas, & Ranganath, 2007; McKenzie et al., 2014), supporting the theory that the elements of the map are bound together within distinct contexts.

Our findings suggest that the brain utilizes the same neural system for representing and navigating continuous space to code the relationship between discrete entities in an abstract space. Further, they suggest that accurate inferences about relative ranks of novel pairs of individuals may depend on the ability to find a direct route in a multidimensional space. This vector-based navigation over the cognitive map may be critical for efficient decision making and knowledge generalization. Moreover, accurate knowledge about the position of others in a social space should provide a solid foundation for sound inferences, thereby supporting effective model-based decision making. We found that the same cognitive map constructed by the HC-EC system is present in other brain areas, including the interconnected vmPFC/mOFC (Barbas & Blatt, 1995; Howard Eichenbaum, 2017; Insausti & Muñoz, 2001; Preston & Eichenbaum, 2013; Wikenheiser et al., 2017) and neighboring central/lateral OFC, generally supporting the theory that OFC represents a cognitive map of task space (Schuck et al., 2016; Wikenheiser & Schoenbaum, 2016; Wilson et al., 2014), though in our study not only of the behaviorally relevant task space, but also the broader task space. Moreover, we show here that the OFC's representation of the task space respects map-like Euclidian distances of vectors through that space and that OFC activity reflects these distances to latent hubs retrieved from memory to guide inference. These findings thus cast light on why the OFC plays a critical role in model-based inference (Jones et al., 2012).

Finally, we suggest that the HC-EC system may play a key role in constructing a global map from local experiences, which may guide model-based decisions in vmPFC/mOFC, a region

previously implicated in value-guided choice (Boorman et al., 2009; Chib, Rangel, Shimojo, & O'Doherty, 2009; Grabenhorst & Rolls, 2011; Hunt et al., 2012; Lim et al., 2011; Papageorgiou et al., 2017; Strait et al., 2014). This same cognitive map appears to further guide how humans integrate knowledge in the social domain, a critical ability for navigating our social worlds.

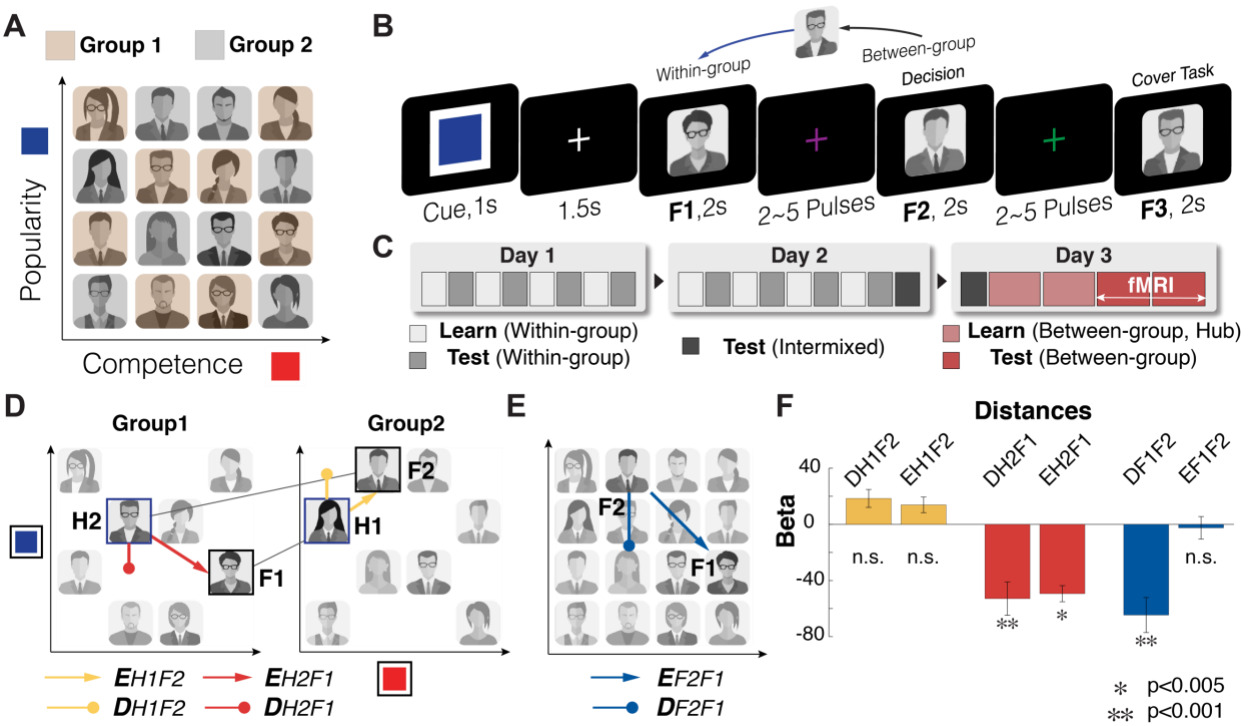

Figure 1. A. Participants learned the rank of members of each of two groups (brown and gray) separately in two dimensions: competence and popularity. Crucially, subjects were never shown the 1- or 2-D structures but could infer them by making transitive inferences. **B.** Illustration of a trial of the fMRI experiment. Participants made inferences about the relative status of a novel pair (F1 and F2) in a given dimension (signaled by Cue color). A cover task (to indicate the gender of the face stimulus, F3) followed at the end of every trial. **C.** On day 1 and day 2, during learning blocks participants learned within-group ranks of the two groups in each of two dimensions through binary decisions about the status of members who differed by only one rank level in a given dimension. Test blocks tested subjects' knowledge of the two 1-D hierarchies that could be constructed using transitive inferences for each group separately. On day3, subjects learned from between-group comparisons limited to 'hub' individuals, which created a unique path between groups per person in each dimension. Subsequently, on day3, participants were asked to infer the unlearned between-group status while undergoing fMRI. **D.** Participants could use hubs to infer the relationship between novel pairs. Possible trajectories for two example inferences can be seen. Two distances are shown for each trajectory: the behaviorally-relevant 1-D distance (D) and the 2-D Euclidian distance (E). Subjects could use either of two trajectories: a forward inference from F1 to its hub (H1) that has a unique connection to F2 (1-D distance: D_{H1F2}, Euclidian distance: E_{H1F2}) which is shown in yellow; or a backward inference from F2 to its hub (H2) that has a unique connection to F1 ($D_{H2F1;}$ E_{H2F1}) which is shown in red. **E.** As alternative paths, subjects may not use the hubs, but instead compute the distance in the relevant dimension between F1 and F2 directly (D_{F1F2}), or their Euclidean distance (E_{F1F2}) in the combined cognitive map of two groups (blue). **F.** Multiple linear regression results show that both the rank distance (D_{H2F1}) and the Euclidean distance from H2 (E_{H2F1}), but not from H1, significantly explain variance in reaction times (RT), in addition to the direct distance between F1 and F2 (D_{F1F2}), while competing with other distance terms.

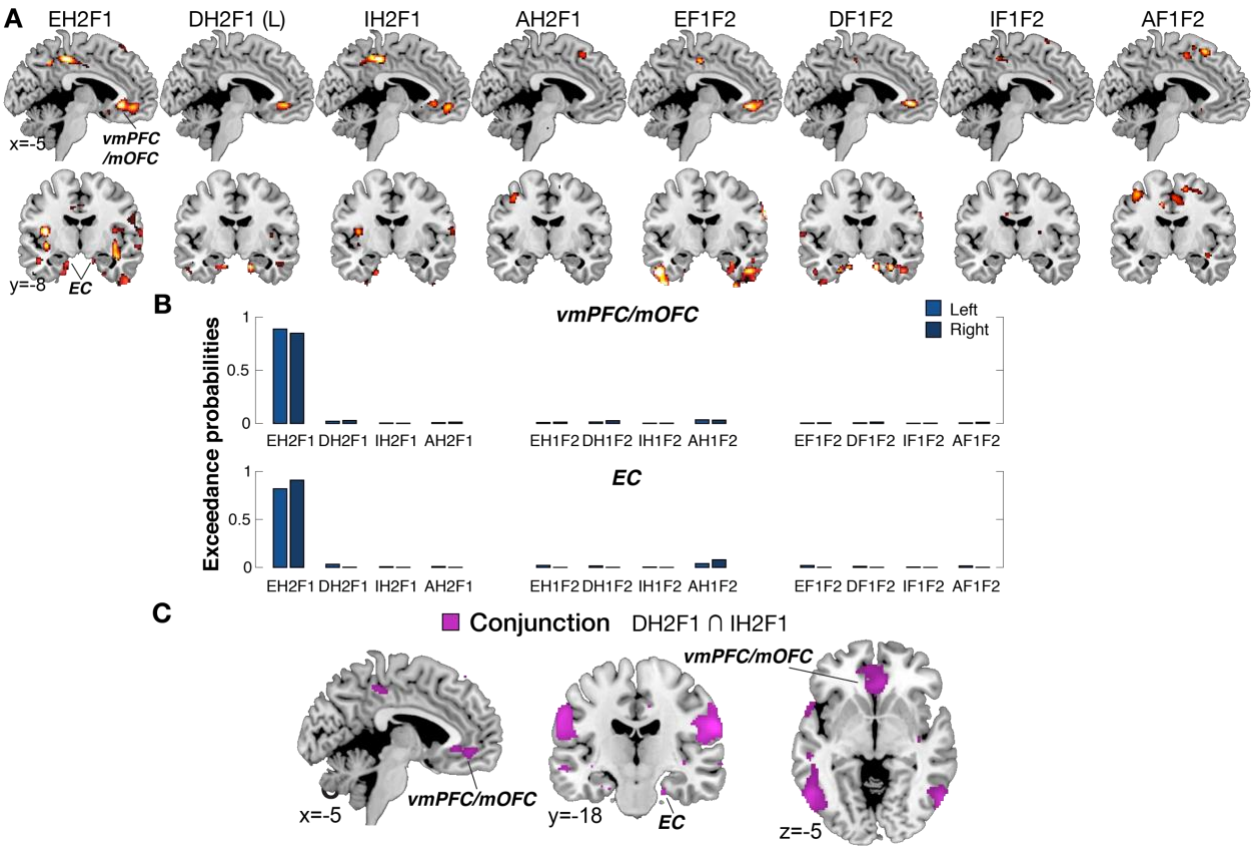

Figure 2. A. The bilateral entorhinal cortex (EC; p_{FWE}<0.05 corrected at peak level within a small volume ROI) and ventromedial prefrontal cortex (vmPFC/mOFC, p<0.05, whole-brain corrected with TFCE) encode the Euclidean distance from the hub inferred from F2 (H2) to F1 in the 2-D social space (E_{H2F1}). Whole-brain parametric analyses showing neural correlates of each of the distance metrics that could theoretically drive inferences between pairs at the time of decisions (F2 presentation). D: 1-D rank distance in the task-relevant dimension (D_{H2F1} and D_{F1F2}); L: the shortest link distance between F1 and F2 (L equals to D_{H2F1}+1); I: the 1-D rank distance in the task-irrelevant dimension (I_{H2F1} and I_{F1F2}); A: the cosine vector angle (A_{H2F1} and A_{F1F2}). **B.** The results of Bayesian model selection (BMS). The exceedance probabilities revealed that the Euclidean distance from the hub (E_{H2F1}) best accounted for variance in both EC and vmPFC/mOFC activity compared to the other distance measures, providing evidence that these regions compute or use a Euclidian distance metric to a retrieved hub (H2) in abstract space in order to infer the relationship between F1 and F2. **C.** The conjunction analysis shown in purple revealed that both D_{H2F1} and I_{H2F1} are both encoded in the right vmPFC/mOFC (p_{SVC}=0.033) and the left vmPFC/mOFC (p_{SVC}=0.054) and the right EC (p_{SVC}=0.067). For visualization purposes, the whole brain maps are thresholded at p<0.005 uncorrected.

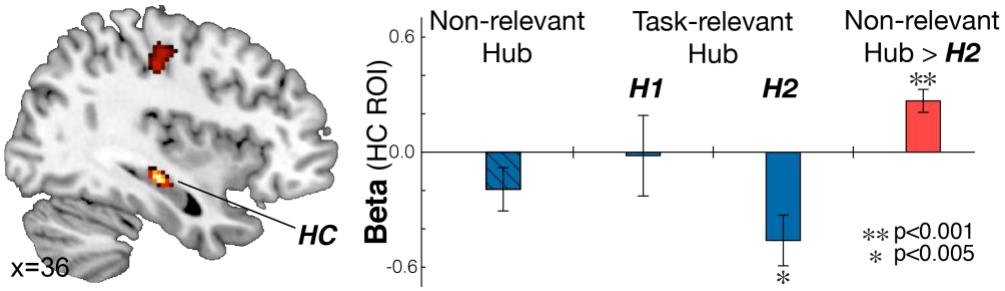

Figure 3. The result of repetition suppression analyses when one of the eight hubs was presented randomly following F2 presentation, as subjects performed a cover task (F3 presentation). Left: BOLD contrast of task-irrelevant hub (Non-hub) > H2, displayed at $p<0.005$ uncorrected (no masking is applied to the image). HC effect is significant at $p_{FWE}=0.03$, small volume corrected at the peak-level, $t_{26}=3.81$, [36,-24,-10]). Right: beta estimates from an independently defined right HC ROI (Yushkevich et al., 2015) (see **Fig. 4C**). Activity in the right hippocampus (HC) was suppressed when the relevant hub (H2) was presented, compared to matched Non-hubs. The activity in the right HC was not suppressed when the hub inferred from F1 (H1) was presented. This result suggests that HC reinstates the neural representation of the specific task-relevant hub (H2) to guide inferences.

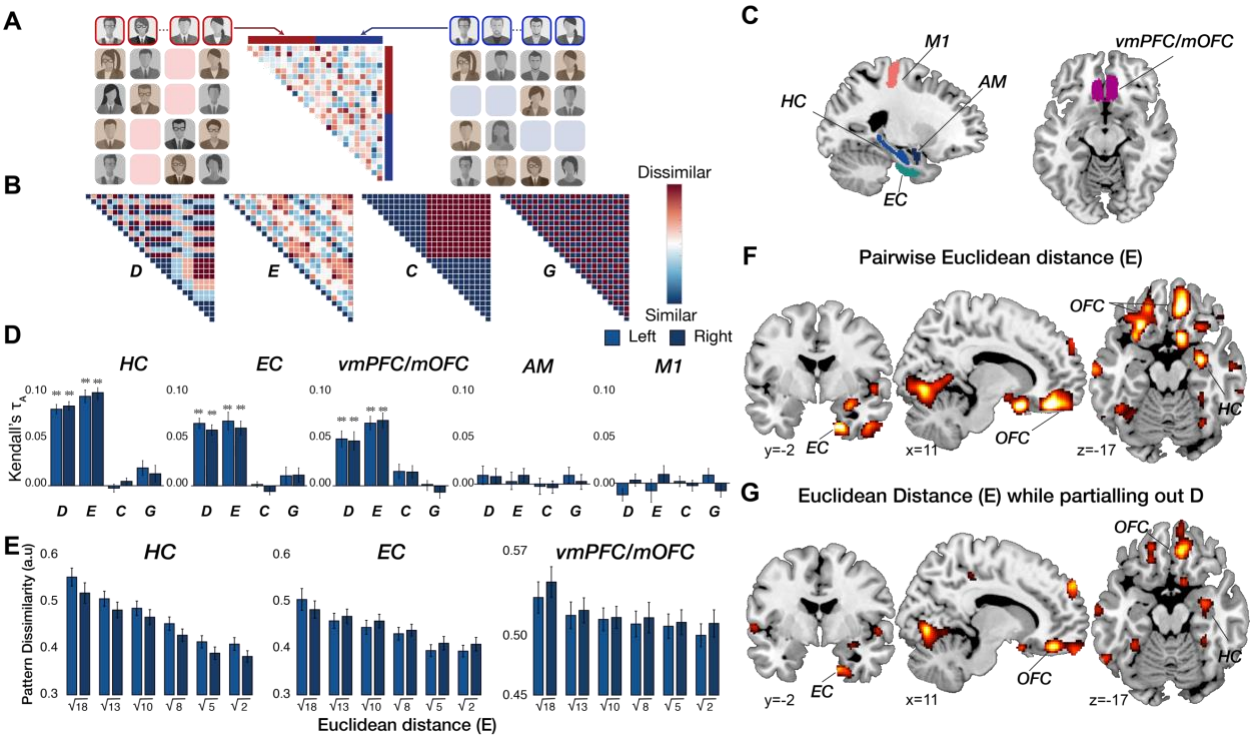

Figure 4. Representational similarity analysis (RSA). **A.** The representational dissimilarity matrix (RDM) was computed in a priori regions of interest (ROIs) from the pairwise Euclidean distance in the multi-voxel activity patterns evoked when face stimuli were presented at the time of F1 and F2. People were modeled separately when they were shown in the competence (left panel) and popularity contexts (right panel). **B.** The neural RDM was explained by model predictions of four separate dissimilarity matrices, including pairwise differences in the rank in the task-relevant dimension (D), pairwise Euclidean distances on the 2-D social space (E), the behavioral context indicating for which social hierarchy dimension the face was presented (C), and in which group (group 1 or 2) the face belonged during training (G). **C.** Regions of interest (ROIs) generated independently from probabilistic maps of other studies (the HC (Yushkevich et al., 2015), EC (Amunts et al., 2005; Zilles & Amunts, 2010), amygdala (AM) (Tzourio-Mazoyer et al., 2002), vmPFC/mOFC (Neubert et al., 2015), and primary motor cortex (MT) (Glasser et al., 2016)). All ROIs were defined bilaterally. **D.** Kendall's τ indicates to what extent a predictor RDM explains the pattern dissimilarity between voxels in each of the ROIs. The model RDMs of D and E, but not C or G, show robust effects on the pattern dissimilarity estimated in the HC, EC, and vmPFC/mOFC (p_{FWE}<0.01, multiple comparisons are corrected with the Holm-Bonferroni method). We did not find any significant effects in the AM or a control region in the M1. **E.** The dissimilarity between brain activity patterns estimated in bilateral HC, EC, and vmPFC/mOFC increases in proportion to the true pairwise Euclidean distance between individuals in the 2-D abstract space. **F.** Whole-brain searchlight RSA indicates effects of E in the HC, EC, mOFC (a part of vmPFC), central OFC, and lateral OFC, among other regions (p_{TFCE}<0.05). **G.** The activity patterns in the HC, EC, and central and medial OFC are still explained by the model RDM for pairwise Euclidean distance (E) after partialling out its correlation with the model RDM for D (p_{TFCE}<0.05; see **Supplementary Fig.7A**). For visualization purposes, the whole-brain searchlight maps are thresholded at p<0.005 uncorrected.

REFERENCES

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., … Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: Intersubject variability and probability maps. In *Anatomy and Embryology* (Vol. 210, pp. 343–352). https://doi.org/10.1007/s00429-005-0025-5

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719–722. https://doi.org/10.1038/nature21692

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433. https://doi.org/10.1038/s41586-018-0102-6

Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. *Neuron*, *102*(5), 1066-1075.e5. https://doi.org/10.1016/j.neuron.2019.03.034

Barbas, H., & Blatt, G. J. (1995). Topographically specific hippocampal projections target functionally distinct prefrontal areas in the rhesus monkey. *Hippocampus*, *5*(6), 511–533. https://doi.org/10.1002/hipo.450050604

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2016). Repetition suppression: A means to index neural representations using BOLD? *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2015.0355

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018, October 24). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*. Cell Press. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

Bickart, K. C., Wright, C. I., Dautoff, R. J., Dickerson, B. C., & Barrett, L. F. (2011). Amygdala volume and social network size in humans. *Nature Neuroscience*, *14*(2), 163–164. https://doi.org/10.1038/nn.2724

Boccara, C. N., Nardin, M., Stella, F., O'Neill, J., & Csicsvari, J. (2019). The entorhinal cognitive map is attracted to goals. *Science*, *363*(6434), 1443–1447. https://doi.org/10.1126/science.aav4837

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green

Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Boorman, E. D., Rajendran, V. G., O'Reilly, J. X., & Behrens, T. E. (2016). Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus. *Neuron*, *89*(6), 1343–1354. https://doi.org/10.1016/j.neuron.2016.02.014

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using Grid Cells for Navigation. *Neuron*, *87*(3), 507–520. https://doi.org/10.1016/j.neuron.2015.07.006

Butler, W. N., Hardcastle, K., & Giocomo, L. M. (2019). Remembered reward locations restructure entorhinal spatial maps. *Science*, *363*(6434), 1447–1452. https://doi.org/10.1126/science.aav5297

Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding Individual Episodic Memory Traces in the Human Hippocampus. *Current Biology*, *20*(6), 544–547. https://doi.org/10.1016/j.cub.2010.01.053

Chan, S. C. Y., Niv, Y., & Norman, K. A. (2016). A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex. *The Journal of Neuroscience*, *36*(30), 7817–7828. https://doi.org/10.1523/jneurosci.0659-16.2016

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *29*(39), 12315–12320. https://doi.org/10.1523/JNEUROSCI.2575-09.2009

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Danjo, T., Toyoizumi, T., & Fujisawa, S. (2018). Spatial representations of self and other in the hippocampus. *Science*, *359*(6372), 213–218. https://doi.org/10.1126/science.aao3898

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386. https://doi.org/10.1016/j.tics.2007.08.001

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661. https://doi.org/10.1038/nature08704

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152.

https://doi.org/10.1146/annurev.neuro.30.051606.094328

Eichenbaum, Howard. (2017, September 29). Prefrontal-hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*. Nature Publishing Group. https://doi.org/10.1038/nrn.2017.74

Eichenbaum, Howard, & Cohen, N. J. (2014). Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? *Neuron*. https://doi.org/10.1016/j.neuron.2014.07.032

Eichenbaum, Howard, Otto, T., & Cohen, N. J. (1996). The hippocampal system: Dissociating its functional components and recombining them in the service of declarative memory. *Behavioral and Brain Sciences*, *19*(4), 772–776. https://doi.org/10.1017/s0140525x00043971

Ekstrom, A. D., & Ranganath, C. (2018, September 1). Space, time, and episodic memory: The hippocampus is all over the cognitive map. *Hippocampus*. John Wiley & Sons, Ltd. https://doi.org/10.1002/hipo.22750

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*. https://doi.org/10.1037/0033-295X.99.4.689

FitzGerald, T. H. B., Seymour, B., & Dolan, R. J. (2009). The Role of Human Orbitofrontal Cortex in Value Comparison for Incommensurable Objects. *Journal of Neuroscience*, *29*(26), 8388–8395. https://doi.org/10.1523/jneurosci.0717-09.2009

Frank, M. J., Rudy, J. W., Levy, W. B., & O'Reilly, R. C. (2005). When logic fails: Implicit transitive inference in humans. *Memory and Cognition*, *33*(4), 742–750. https://doi.org/10.3758/BF03195340

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *NeuroImage*, *25*(3), 661–667. https://doi.org/10.1016/j.neuroimage.2005.01.013

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*. https://doi.org/10.7554/eLife.17086

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., … Van Essen, D. C. (2016). The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, *19*(9), 1175–1187. https://doi.org/10.1038/nn.4361

Grabenhorst, F., & Rolls, E. T. (2011, February 1). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*. Elsevier Current Trends. https://doi.org/10.1016/j.tics.2010.12.004

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.

https://doi.org/10.1016/j.tics.2005.11.006

Guise, K. G., & Shapiro, M. L. (2017). Medial Prefrontal Cortex Reduces Memory Interference by Modifying Hippocampal Encoding. *Neuron*, *94*(1), 183-192.e8. https://doi.org/10.1016/j.neuron.2017.03.011

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Howard, L. R., Javadi, A. H., Yu, Y., Mill, R. D., Morrison, L. C., Knight, R., … Spiers, H. J. (2014). The Hippocampus and Entorhinal Cortex Encode the Path and Euclidean Distances to Goals during Navigation. *Current Biology*, *24*(12), 1331–1340. https://doi.org/10.1016/J.CUB.2014.05.001

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, *15*(3), 470–476. https://doi.org/10.1038/nn.3017

Insausti, R., & Muñoz, M. (2001). Cortical projections of the non-entorhinal hippocampal formation in the cynomolgus monkey (Macaca fascicularis). *European Journal of Neuroscience*, *14*(3), 435–451. https://doi.org/10.1046/j.0953-816X.2001.01662.x

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, *338*(6109), 953–956. https://doi.org/10.1126/science.1227489

Kaplan, R., & Friston, K. J. (2019). Entorhinal transformations in abstract frames of reference. *PLoS Biology*, *17*(5), e3000230. https://doi.org/10.1371/journal.pbio.3000230

Klein-Flugge, M. C., Barron, H. C., Brodersen, K. H., Dolan, R. J., & Behrens, T. E. J. (2013). Segregated Encoding of Reward-Identity and Stimulus-Reward Associations in Human Orbitofrontal Cortex. *Journal of Neuroscience*, *33*(7), 3202–3211. https://doi.org/10.1523/JNEUROSCI.2532-12.2013

Koster, R., Chadwick, M. J., Chen, Y., Berron, D., Banino, A., Düzel, E., … Kumaran, D. (2018). Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes. *Neuron*, *99*(6), 1342-1354.e6. https://doi.org/10.1016/J.NEURON.2018.08.009

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Kropff, E., Carmichael, J. E., Moser, M. B., & Moser, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature*, *523*(7561), 419–424. https://doi.org/10.1038/nature14622

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations Underlying Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing Self-Relevant Information. *Neuron*, *92*(5), 1135–1147. https://doi.org/10.1016/j.neuron.2016.10.052

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Kumaran, D., Melo, H. L., & Duzel, E. (2012). The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. *Neuron*, *76*(3), 653–666. https://doi.org/10.1016/j.neuron.2012.09.035

Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast Sequences of Non-spatial State Representations in Humans. *Neuron*, *91*(1), 194–204. https://doi.org/10.1016/j.neuron.2016.05.028

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. *Journal of Neuroscience*, *31*(37), 13214–13223. https://doi.org/10.1523/jneurosci.1246-11.2011

McKenzie, S., Frank, A. J., Kinsky, N. R., Porter, B., Rivière, P. D., & Eichenbaum, H. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*, *83*(1), 202–215. https://doi.org/10.1016/j.neuron.2014.05.019

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., … Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, *26*(4), 490–503. https://doi.org/10.1016/j.mri.2007.08.006

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. *Annual Review of Neuroscience*, *31*(1), 69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723

Nau, M., Navarro Schröder, T., Bellmund, J. L. S., & Doeller, C. F. (2018). Hexadirectional

coding of visual space in human entorhinal cortex. *Nature Neuroscience*, *21*(2), 188–190. https://doi.org/10.1038/s41593-017-0050-8

Neubert, F.-X., Mars, R. B., Sallet, J., & Rushworth, M. F. S. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), 1–10. https://doi.org/10.1073/pnas.1410767112

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron*, *75*(6), 1114–1121. https://doi.org/10.1016/j.neuron.2012.07.023

Nili, H., Walther, A., Alink, A., & Kriegeskorte, N. (2016). Inferring exemplar discriminability in brain representations. *Manuscript in Preparation*, 080580. https://doi.org/10.1101/080580

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Noonan, M. A. P., Sallet, J., Mars, R. B., Neubert, F. X., O'Reilly, J. X., Andersson, J. L., … Rushworth, M. F. S. S. (2014). A Neural Circuit Covarying with Social Hierarchy in Macaques. *PLoS Biology*, *12*(9), e1001940. https://doi.org/10.1371/journal.pbio.1001940

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press.

Omer, D. B., Maimon, S. R., Las, L., & Ulanovsky, N. (2018). Social place-cells in the bat hippocampus. *Science*, *359*(6372), 218–224. https://doi.org/10.1126/science.aao3474

Papageorgiou, G. K., Sallet, J., Wittmann, M. K., Chau, B. K. H., Schüffelgen, U., Buckley, M. J., & Rushworth, M. F. S. (2017). Inverted activity patterns in ventromedial prefrontal cortex during value-guided decision-making in a less-is-more task. *Nature Communications*, *8*(1), 1886. https://doi.org/10.1038/s41467-017-01833-5

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014

Preston, A. R., & Eichenbaum, H. (2013, September 9). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*. Cell Press. https://doi.org/10.1016/j.cub.2013.05.041

Rushworth, M. F. S., Noonan, M. A. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011,

June 23). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*. Cell Press. https://doi.org/10.1016/j.neuron.2011.05.014

Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., … Rushworth, M. F. S. (2011). Social network size affects neural circuits in macaques. *Science*, *334*(6056), 697–700. https://doi.org/10.1126/science.1210027

Schiller, D., Eichenbaum, H., Buffalo, E. A., Davachi, L., Foster, D. J., Leutgeb, S., & Ranganath, C. (2015). Memory and Space: Towards an Understanding of the Cognitive Map. *Journal of Neuroscience*, *35*(41), 13904–13911. https://doi.org/10.1523/JNEUROSCI.2618-15.2015

Schlichting, M. L., & Preston, A. R. (2014). Memory reactivation during rest supports upcoming learning of related content. *Proceedings of the National Academy of Sciences*, *111*(44), 15845–15850. https://doi.org/10.1073/pnas.1404396111

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/J.NEURON.2016.08.019

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Stalnaker, T. a, Cooch, N. K., & Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nature Neuroscience*, *18*(5), 620–627. https://doi.org/10.1038/nn.3982

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/J.NEUROIMAGE.2009.03.025

Strait, C. E., Blanchard, T. C., & Hayden, B. Y. (2014). Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. *Neuron*, *82*(6), 1357–1366. https://doi.org/10.1016/j.neuron.2014.04.032

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods*. https://doi.org/10.3758/s13428-015-0641-9

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron*, *95*(6), 1395-1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025

Tang, E., Mattar, M. G., Giusti, C., Lydon-Staley, D. M., Thompson-Schill, S. L., & Bassett, D. S. (2019). Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, *22*(6), 1000–1009. https://doi.org/10.1038/s41593-019-0400-9

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A Map for Social Navigation in the Human Brain. *Neuron*, *87*(1), 231–243. https://doi.org/10.1016/j.neuron.2015.06.011

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology*, *29*(7), 1226-1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tompary, A., & Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. *Neuron*, *96*(1), 228-241.e5. https://doi.org/10.1016/J.NEURON.2017.09.005

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., … Daw, N. D. (2019). Hippocampal Contributions to Model-Based Planning and Spatial Memory. *Neuron*. https://doi.org/10.1016/j.neuron.2019.02.014

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. S. (2010). Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, *65*(6), 927–939. https://doi.org/10.1016/j.neuron.2010.02.027

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., … Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Weiskopf, N., Hutton, C., Josephs, O., & Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3 T

and 1.5 T. *NeuroImage*, *33*(2), 493–504. https://doi.org/10.1016/j.neuroimage.2006.07.029

Whittington, J. C. R., Muller, T. H., Mark, S., Barry, C., & Behrens, T. E. J. (2018). Generalisation of structural knowledge in the hippocampal-entorhinal system. *Advances in Neural Information Processing Systems*, 8484–8495.

Wikenheiser, A. M., Marrero-Garcia, Y., & Schoenbaum, G. (2017). Suppression of Ventral Hippocampal Output Impairs Integrated Orbitofrontal Encoding of Task Structure. *Neuron*, *95*(5), 1197-1207.e3. https://doi.org/10.1016/j.neuron.2017.08.003

Wikenheiser, A. M., & Schoenbaum, G. (2016, August 3). Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*. Nature Publishing Group. https://doi.org/10.1038/nrn.2016.56

Wilson, R. C., & Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004237

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273. https://doi.org/10.1126/science.1223252

Yushkevich, P. A., Amaral, R. S. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., … Zeineh, M. M. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *NeuroImage*, *111*, 526–541. https://doi.org/10.1016/j.neuroimage.2015.01.004

Zilles, K., & Amunts, K. (2010, February 4). Centenary of Brodmann's map conception and fate. *Nature Reviews Neuroscience*. Nature Publishing Group. https://doi.org/10.1038/nrn2776

METHODS

Participants

A total of 33 participants (16 female, age range: 19–23, normal or corrected to normal vision) were recruited for this study via the University of California, Davis online recruitment system. Six participants were excluded due to strong head movements larger than the voxel size of 3 mm. In total, 27 participants entered the analysis (mean age: 19.37±0.26, standard error mean (SEM)). The study was approved by the local ethics committee, all relevant ethical regulations were followed, and participants gave written consent before the experiment.

Stimulus

The stimuli consisted of 16 grayscale photographic images of faces (Strohminger et al., 2016) and two colored cues (red and blue squares). Each of the colored cues indicates the task-relevant dimension of social hierarchy for the current trial. The red square indicated the competence hierarchy for one-half of participants and the popularity hierarchy for the other half. All images were adjusted to the same mean grayscale value. The inter-trials fixation target was a white cross in the middle of a black screen. For hub learning and fMRI experiment, the inter-stimuli fixation target was a purple cross (between F1 and F2) and a green cross (between F2 and F3) in the middle of a black screen, which informs the progress of each trial to participants. The stimuli were presented to participants through a mirror mounted on the head coil. Note that face stimuli presented in this paper are license-free images for display purposes. Prior to the experiment on the first day of training, participants performed a 1-back task where they viewed each individual face three times to minimize stimulus novelty effects.

Social hierarchies

Participants were asked to learn the relative ranks of 16 individuals (face stimuli) in two dimensional social hierarchies–popularity and competence. The 16 face stimuli were introduced as entrepreneurs; participants were asked to learn about which individuals were more capable to attract crowd funds (popularity) and which individuals had higher technical proficiency (competence) and used this information to guide investment decisions.

Each hierarchy has four levels of ranks. Four individuals were allocated at the same rank at each level of the hierarchy. Therefore, the structure of multidimensional social hierarchies is 4×4 (**Fig. 1A**). The rank of an individual in one dimension was not related to his/her rank in the other dimension. For instance, the rank of four individuals who are at the 1st rank in the popularity

hierarchy are the 1st, the 2nd, the 3rd, and the 4th, respectively in the competence hierarchy. During first two days of training, the relative status of one individual is only compared to one-half of the other face stimuli, which implicitly creates two groups in which each group was comprised of eight individuals (**Supplementary Fig.1**). In each group, two individuals were allocated into the same rank of each of the hierarchies. The sub-group structure is shown in **Fig. 1A**. The allocation of each face to the position in the social hierarchies was pseudo-randomized, to make sure that any of visual features of the face (gender, race, and age) was not associated with the rank group of the individuals. To do this, we prepared eight stimuli sets. Each of the stimuli set comprises of different 16 faces. A stimuli set was randomly assigned across participants.

Task instruction and experiment procedures.

Participants were instructed to imagine that they were a venture capitalist and decide where to invest after learning the relative ranks of 16 entrepreneurs in two independent dimensions − competence and popularity. Participants were asked to learn which individual was better in technical proficiency – competence hierarchy, and which individual was better in attracting crowd funding – popularity hierarchy. During the learning block, participants were presented two face stimuli of entrepreneurs with the contextual cue of task-relevant hierarchy, chose one who is superior to the other (the higher rank individual) in the given dimension, and received feedbacks at the end of every trial. Participants were told that they would need to use the knowledge acquired during the learning block to decide which entrepreneur they want to invest in which the ability in one social hierarchy dimension is important. During the test block, participant chose one of two face stimuli who they believed was higher in the given social hierarchy. They did not receive any feedback during test block.

Before training, the following instructions were clearly given to participants: (1) two entrepreneurs presented in the learning block have one rank difference whereas two entrepreneurs presented in the test block have one or more rank differences. (2) Multiple individuals could be allocated to the same rank. Importantly, participants have never been given any information implying the structure of social hierarchies − such as the total number of ranks in each of social hierarchies, and a number of individuals allocated into the same rank. Each subject participated in behavioral training across three separate days, at least 48 hours apart. After the behavior training on the third day, they participated in the fMRI experiment.

Behavioral training

The behavioral training was comprised by the 'learning' block and the 'test' block (**Fig. 1C**). In the beginning of each of the mini-blocks, participants were presented which block they were in. The purpose of the learning block paradigm was to provide an experimental setting in which participants would gradually acquire robust knowledge of social hierarchies of two groups of individuals and use this to make successful transitive inferences during trials in the test block.

For the test block during the training, participants were asked to infer the relative rank between two individuals. To make a correct inference, therefore, participants need to use successful transitive inference for the test block. If they adopted an alternative cognitive process such as statistical learning – comparing the values assigned to each of the face stimuli according to their number of wins/loses −, participants could not distinguish the second and third rank individuals since their number of wins/loses were equal. Considering that, participants who successfully distinguished the second and third rank individuals above chance while reaching above 85% accuracy in each test block were invited to participate in the following procedure of the experiment.

It is important to note that, during three days of training, each of the 16 individuals has been exposed with the same numbers of times to the participants. For the trial in which participants could not respond within 2s, feedback was not given. This missing trial was tested again after a random number of trials to ensure all participants have the same level of knowledge. Participants were never asked to combine individuals' ranks in both dimensions to make decisions. Participants have never seen either one-dimensional (1-D) or two-dimensional (2-D) global structure of social hierarchies.

Learning relative ranks of within-group members (The learning block of day 1 and day 2 training)

During the learning block, participants were presented two face stimuli on a black screen having one rank difference with a colored contextual cue which indicated which was the task-relevant social hierarchy in the current trial (learning block in **Supplementary Fig.1A**). They were asked to indicate who was superior to the other in the given social hierarchy. Participants learned the relative status of all possible one rank difference pairs with feedback. The feedback (correct/ incorrect) was followed at the end of all responded trials. For the learning block of day 1 training, participants learned the relative status of an eight-individual group in one of two social hierarchies for the first mini-block (e.g. the hierarchy in competence dimension for group 1 individuals). For the second mini-block, they learned the relative status of the other eight individual group in the other social hierarchy (e.g. the hierarchy in popularity dimension for group 2 individuals). For the learning block of day 2 training (the learning block in **Supplementary Fig.1B**), they learned the

relative status of each group individuals in the unlearned hierarchy dimension (e.g. the hierarchy in popularity dimension for group1 individuals and the hierarchy in competence dimension for group2 individuals). After two days of training, participants were able to develop the knowledge about two different social hierarchies of individuals in two groups. The right panel in **Supplementary Fig.1** shows the possible structure of the cognitive map that participants could build at the end of each training day. For each of day 1 and day 2 training, participants complete eight mini-blocks of learning block (**Fig 1C**). One-half of participants learned the relative ranks of group 1 in the competence dimension for the first day and the other half of participants learned the relative ranks of the group 1 in the popularity dimension for the first day.

Inferences of relative ranks in one-dimensional social hierarchy (The test block of day1 and day2)

After the learning block, we tested whether participants could generalize their knowledge to infer the relative status between individuals having one or more ranks difference. This test block was followed by each mini-block of the learning blocks (the test block in **Supplementary Fig.1**). During the test block, all possible pairs of within-group individuals were presented to participants except for the pair of individuals who are at the same rank in the given dimension. In each trial, participants were asked to indicate a face who was superior to the other in the given dimension. Participants were instructed that their choices would count towards their final payout. No feedback was given during test block to prevent further learning.

Flexible inferences about relative ranks in intermixed contexts (The test 2 block in day2)

At the end of the second-day training, an additional test block block was followed. During this test 2 block, the trials asking the relative rank of group 1 individuals and the trials asking the relative rank of group 2 were intermixed while the task-relevant dimension was also intermixed across trials (the test 2 block in **Fig S2B**). Consistent with the previous blocks, participants were asked to indicate one of two face stimuli who is superior to the other in the given dimension. No feedback was given during test 2 block. Participants were instructed that their choices would count towards their final payout.

Learning relative ranks of between-group with 'hub' (The learning block in day 3)

On the third-day training, participants learned the relative ranks of pairs of between-group individuals for the first time. Importantly, the purpose of the learning block paradigm in the third-day training was to provide a limited knowledge to participants about relative ranks between-

group individuals. That is, participants did not learn the relative rank of all pairs of between-group individuals but only the relative rank of the selected pairs of between-group individuals.

During the hub learning block, participants learned the relative status between one individual in one group (hub) and another individual in the other group (non-hub) who has one rank difference in the given dimension. In each group, four individuals (two per each group) were selected as hubs in one dimension. In the other dimension, different four individuals played a role as hubs (eight hubs in total; **Supplementary Fig.2E**). For those two hubs in each group, one was at the second rank, and the other was at the third rank in the given dimension. Each of the hubs was paired with four different individuals in the other group. With this procedure, all eight individuals in one group (non-hub) were paired with two selected individuals in the other group in one dimension (they were never paired with the other six individuals who were not selected as hubs). In particular, a hub in group 1 who was at the third-rank in the dimension was paired with four individuals in group 2 including two second-rank individuals and two fourth-rank individuals (the top panel in **Supplementary Fig.2B**). The other hub in group 1 who was at the second-rank in the given dimension was paired with the other four individuals in group 2 including two first-rank individuals and two third-rank individuals (the bottom panel in **Supplementary Fig.2B**). This is also true for hubs in group 2 (**Supplementary Fig.2C**). In doing so, participants have learned the relative rank of some pairs of between-group individuals who have one rank difference each other as they learned for the pairs of within-group individuals during the previous learning block. The hub learning block paradigm eventually allows us to create a unique path between members in different groups. That is, each of 12 non-hubs individuals (six per each group; **Supplementary Fig.2D**) has a unique connection to a specific hub in the other group (one among eight hubs in **Supplementary Fig.2E**) in one of two hierarchy dimensions. Note that the hubs in competence dimension differed from the hubs in popularity dimension.

For each trial in the hub learning block, three face stimuli (F1, F2, and F3) were presented for 2 s sequentially after the presentation of a conditional cue (1 s) indicating the task-relevant dimension of the current trial (**Supplementary Fig.2A**). Participants were asked to indicate one who is superior to the other between F1 and F2 in the given dimension while F2 was presented. The feedback (correct/ incorrect) was followed after each decision (2 s). Between F1 and F2, one was the hub in the given dimension and the other was a non-hub individual in the different group who has one rank difference from the hub. While presenting F3 (2 s) at the end of every trial, participants were asked to press a button according to the gender of the F3 face stimuli. F3 was selected from 12 non-hubs in the given dimension (**Supplementary Fig.2D**). No feedback was given for the gender discrimination task. Inter-stimuli interval (ISI) was 2 s. Inter-trials interval (ITI)

was 4 s. While learning between-group relationship via hubs, participants became familiar with the task that we used for fMRI experiment (**Fig. 1B**). The number of presentations of each of the face stimulus was controlled to be equivalent.

fMRI experiment

Inferences about the relative ranks of novel pairs (The test block in day3)

The purpose of the fMRI experiment paradigm was to test whether and how participants represent their knowledge of social hierarchies of the two groups of individuals and make a successful inference about relative ranks of novel pairs of individuals. **Fig. 1B** illustrates the procedure of an example trial of the fMRI experiment. In each trial, three face stimuli (F1, F2, and F3) were shown sequentially following a conditional cue (1 s) with an inter-stimuli fixation cross (1.5 s). The color of a square shown in conditional cue indicated the relevant dimension of the current trial. The same number of trials were shown per dimension of social hierarchies in a block. The trials were intermixed in random order. Each face stimuli were shown for 2 s. Between face stimuli, we presented a fixation cross for inter-stimuli-intervals (ISI) which were jittered between 2 ~ 5 pulses (TR=1200ms). The first decision was made during the F2 presentation. Participants were asked to press a button to indicate who is superior to the other between F1 and F2 in the given social hierarchy dimension. They were asked to respond as fast as possible but also as accurate as possible. No feedback was given. The second decision was made during the F3 presentation. Participants were asked to press a button to indicate the gender of F3 as fast as possible. The buttons allocated to indicate the gender of presenting face stimuli were counterbalanced across participants. If the response is missed in the inference decision, we showed 'missed' sign and proceeded to the next trial. The missed trial was tested again after a random number of trials, which allowed us to collect all responses from participants.

The following were not informed to participants: (1) during the fMRI experiment, F1 and F2 were selected from different groups among 12 non-hubs individuals in the given dimension (**Supplementary Fig.2D**) – F1 was selected from group 1 for one-half of trials and F2 was selected from group 1 for the other half –; (2) F3 was selected among eight individuals who played a role in hubs regardless of the social hierarchy dimension (**Supplementary Fig.2E**). All eight hubs were shown the same number of trials at the time of the F3 presentation. Participants were asked to make the same type of decisions continuously as they did for the third-day behavioral training (i.e. choosing a higher hierarchy individual between first two faces in the given context dimension and indicating the gender of the third face). By doing this, unbeknownst to participants,

we were able to test whether and how participants make inferences about the relative position of unlearned pairs of individuals. The fMRI experiment comprised of two blocks. Each block included 104 trials which include all possible between-group pairs of non-hubs who have different ranks in the given dimension. Note that, during the fMRI experiment, all F1-F2 pairs were also presented in reverse order in both context dimensions. The order of the trials was randomized across participants.

Inferences of relative status between unexperienced individuals via hubs

During training, participants never directly learned the relative status between two face stimuli (F1 and F2) presented during the fMRI paradigm. Instead, participants could make transitive inferences about relative status of unlearned pairs via one of two hub individuals (H1 and H2), (**Fig. 1D**). Note that, for every F1-F2 pair, there were only two individuals (H1 and H2) that have been paired with both F1 and F2 during training in the given dimension. That is, H1 belonged to the same group with F2 (within-group) which had uniquely paired with F1 during the hub learning block (between-group). Likewise, H2 belonged to the same group with F1 (within-group) which had uniquely paired with F2 during the hub learning block (between-group). The direction and the distance of inference trajectories on the social cognitive map were, therefore, determined by which of the hub (between H1 and H2) was preferentially recalled by participants for making transitive inferences. The between-group relationship to the hub (F1 → H1 and F2 → H2) had one rank difference in the given dimension. If participants recalled H1, the transitive inference depended on the within-group distance (H1 → F2), and the inference was made in the forward direction (F1 → F2). If participants recalled H2, the transitive inference depended on the within-group distance (H2 → F1), and the inference was made in the backward direction (F2→ F1). We examined which trajectory participants preferentially have chosen for making transitive inferences by examining which unseen hub was selectively retrieved during inferences.

Behavioral data analysis

We estimated the reaction times (RT) and accuracy in inferences of the relative status between a novel pair of individuals (F1 and F2). The RT was measured from the F2 onset to the response. To make successful inferences, participants need to use the cognitive map of social space with the transitive inference via an unseen hub. The inference trajectories, therefore, grounded by the location of the hub. To examine which hub was preferentially selected for inferences, we regressed choice RT on different distance measures of putative inference trajectories using a multiple linear regression model. The accuracy level was not regressed on the distances of

putative trajectories since all participants reached high accuracy (**Supplementary Fig.3B**). As regressors, we included both distances which were measured from each of two potential hubs: the distance between H1 and F2 and the distance between H2 and F1 in addition to the distance between F1 and F2 by allowing them to compete to explain RT variance. Moreover, the distance was measured in both of the rank difference in the task-relevant hierarchy (D) and Euclidean distance (E). Taken together, we regressed RT in inference decisions on four different distance measures of inference trajectories via hubs, $D_{H1F2}, D_{H2F1}, E_{H1F2}$, and E_{H2F1}(**Fig. 1 D**) and two distances measures between F1 and F2, D_{F1F2} and E_{F1F2} (**Fig. 1 E**), (**Eq. 1**).

$$RT = C + \beta_1 E_{H1F2} + \beta_2 E_{H2F1} + \beta_3 E_{F1F2} + \beta_4 D_{H1F2} + \beta_5 D_{H2F1} + \beta_6 D_{F1F2}$$

Eq. 1

The correlation between different distance measures is shown in **Supplementary Fig.5B**. At last, the group level effects of each of the distance measures were tested with a one-sample t-test to account subjects as a random variable.

Functional imaging acquisition

We acquired T2-weighted functional images on a Siemens Skyra 3 Tesla scanner. We used gradient-echo-planar imaging (EPI) pulse sequence that sets the slice angle of 30° relative to the anterior-posterior commissure line, minimizing the signal loss in the orbitofrontal cortex region (Weiskopf, Hutton, Josephs, & Deichmann, 2006). We acquired 38 slices, 3mm thick with the following parameters: repetition time (TR) = 1200 ms, echo time (TE) = 24 ms, flip angle = 67°, field of view (FoV) = 192mm, voxel size = 3 x 3 x 3 mm3. Contiguous slices were acquired in interleaved order. We also acquired a field map to correct for potential deformations with dual echo-time images covering the whole brain, with the following parameters: TR = 630 ms, TE1 = 10 ms, TE2 = 12.46 ms, flip angle = 40°, FoV = 192mm, voxel size = 3 x 3 x 3 mm3. For accurate registration of the EPIs to the standard space, we acquired a T1-weighted structural image using a magnetization-prepared rapid gradient echo sequence (MPRAGE) with the following parameters: TR = 1800 ms, TE = 2.96 ms, flip angle = 7°, FoV = 256mm, voxel size = 1 x 1 x 1 mm3.

Pre-processing

The preprocessing of functional imaging data was performed using SPM12 (Wellcome Trust Centre for Neuroimaging). Images were corrected for slice timing, realigned to the first volume, and realigned to correct for motion using a six-parameter rigid body transformation.

Inhomogeneities created using the phase of nonEPI gradient echo images at 2 echo times were coregistered with structural maps. Images were then spatially normalized by warping subject-specific images to the reference brain on an MNI (Montreal Neurological Institute) coordinate (2mm isotropic voxels). For the univariate analysis images were smoothed using an 8-mm full-width at half maximum Gaussian kernel (Mikl et al., 2008).

Univariate analysis

We implemented a general linear model (GLM) to analyze the fMRI data. The GLM contained separate onset regressors for the contextual cue which indicates the task-relevant dimension, F1, F2, and F3 stimuli presentations for each of the trials. Specifically, the F3 onsets were separately modeled when F3 was (1) the task-relevant hub, H1, (2) the task-relevant hub, H2, and (3) neither of H1 nor H2, but the hub for other pairs of individuals (non-relevant hub). The BOLD signal of the brain was modeled by a stick function for the contextual cue and the F3 presentation and a 2 s boxcar function for the presentation of F1 and F2. The F1 onset regressors were modulated with parametric regressors of the rank of the individual in the task-relevant hierarchy ($F1_R$) and the rank in the task-irrelevant hierarchy ($F1_I$). The F2 onset regressors were modulated by the rank in the task-relevant hierarchy ($F2_R$), the rank in the task-irrelevant hierarchy ($F2_I$), and additional regressors representing the putative inference trajectory which vary according to which structure of cognitive map was tested (**Supplementary Fig.5A**). The onset of button-press was also modeled with a stick function. The 6 motion parameters obtained during realignment were entered into the GLM as a regressor of no interest. The orthogonalization function was turned off. All these regressors were convolved with the canonical hemodynamic response function.

To test whether the brain encodes the trajectories via hubs over Euclidean space, for GLM1, we included parametric regressors of Euclidean distance and cosine angle of the vector between F1 and H2 and those of the vector between F2 and H1 (E_{H2F1}, A_{H2F1}, E_{H1F2}, and A_{H1F2}). The Euclidean distance between face stimuli was defined over the two-dimensional (2-D) space characterized by their relative rank in each of two social hierarchies. The cosine angle represents the normalized function of competence modulated by popularity. The value of these regressors was invariant to the dimension related to the current task.

From the first-level analysis, contrast images of the brain activity correlate with regressors of inference trajectories (E_{H2F1}, A_{H2F1}, E_{H1F2}, and A_{H1F2}) at the time of F2 presentation were estimated from each of participants. Moreover, during the cover task (at the time of F3 presentation), following contrasts images were estimated for the cross stimuli suppression (CSS)

analysis: the brain activity which was suppressed more when F3 was the task-relevant hub, H1 compared to when F3 was non-relevant hubs (H1 < Non-relevant hub); the brain activity which was suppressed more when F3 was the task-relevant hub (H2) compared to when F3 was non-relevant hubs (H2 < Non-relevant hub). The individual contrast images were estimated at the threshold, $p<0.001$, Fisher's Z transformed, and enter into the second-level analysis. We reported the whole-brain threshold-free cluster enhancement (TFCE) corrected signals at the threshold $p_{TFCE}<0.05$ (Smith & Nichols, 2009). Considering that the hippocampus (HC) and entorhinal cortex (EC) were our *a priori* ROI while the EC is prone to signal loss with fMRI, we reported our results in HC and EC at a cluster-defining statistical threshold of $p<0.001$ uncorrected, combined with small volume correction (SVC) in anatomically defined ROIs for correction of multiple comparisons ($p_{FWE}<0.05$ at peak level).

With the GLM2, we tested whether the brain uses different cognitive maps for making inferences in different contextual dimensions. We inputted the rank difference in the task-relevant dimension (D) and that of the task-irrelevant dimension (I) as the parametric regressors which includes the 1-D distances between H2 and F1 and those of H1 and F2 (D_{H2F1}, I_{H2F1}, D_{H1F2} and I_{H1F2}) in addition to the other regressors not associated with the distance measures that we inputted in GLM1. The value of these regressors was dependent on which was the task-relevant dimension. To examine the brain area encodes both D and I, we performed a conjunction analysis in which the individual contrasts images estimated at $p<0.001$ were entered to the second-level analysis (One-way ANOVA). The whole-brain FWE corrected at cluster level ($p_{FWE}<0.05$). With the GLM3, we tested whether the brain has already integrated the cognitive map for group 1 and that for group 2 into a combined cognitive map and encodes the inference trajectories between F1 and F2. We included the regressors of Euclidean distance and cosine angle of the vector between F1 and F2 (E_{F1F2} and A_{F1F2}) in addition to the other regressors that we inputted in GLM1. With the GLM4, we tested whether the brain uses different combined cognitive maps for making inferences in different contextual dimensions. We inputted the rank difference in the task-relevant dimension and that of the task-irrelevant dimension as 1-D distances between F1 and F2 (D_{F1F2} and I_{F1F2}) in addition to the other regressors that we inputted in GLM1. **Supplementary Fig.5A** illustrates the regressors of different models to examine how the brain constructs and use a cognitive map of abstract social hierarchies.

Cross stimuli suppression analysis

Recent findings have shown that the blood-oxygen-level-dependent (BOLD) suppression can be measured not only to repetition of a stimulus, but also to pairs of stimuli that have been

well-learned though association (Barron et al., 2016; Boorman et al., 2016; Grill-Spector et al., 2006; Klein-Flugge et al., 2013). This cross-stimulus suppression (CSS) allows us to examine the underlying neural representations of associative memories. In the current study, if the relevant hub is presented during the suppression phase, at the time of F3 presentation, directly after participants recall the relevant hub for making inferences of relative ranks between faces, then the BOLD signal in the areas encoding the relevant hub should be suppressed compared to the non-relevant hubs. Considering that effects of CSS did not depend on the responses of participants during the cover task, we included the BOLD responses in every F3 presentation into the analysis. Moreover, the BOLD signal should be suppressed specific to the relevant and preferentially selected hub compared to the relevant but unselected hub. Considering that the hippocampus (HC) and entorhinal cortex (EC) system were our *a priori* ROI, we reported our results at a cluster-defining statistical threshold of p<0.001 uncorrected, combined with small volume correction (SVC) in anatomically defined ROIs for correction of multiple comparisons (p_{FWE}<0.05).

Neural model comparison

The different hypothetical structure of the cognitive map and putative inference trajectories cannot be tested in a GLM while there is potential multicollinearity between different distance measures. The cross-correlation between different distance measures is shown in **Supplementary Fig.5B**. To formally compare the predictability of each of distance measures in different models, we used Bayesian model selection (BMS), (Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

We tested whether the brain activity in the vmPFC/mOFC and EC which correlate with the Euclidean distance of the inference trajectory from the hub (E_{H2F1}) is better explained with other alternative distance measures of inference trajectories. To test this, we ran several GLMs in which the brain activity at the time of inferences (F2 presentation) was modeled with only one of candidate distance measures as parametric regressors. Specifically, we compared the models having one of the following distance measures as parametric regressors: E_{H2F1}, E_{H1F2}, E_{F1F2}, D_{H2F1}, D_{H1F2}, D_{F1F2}, A_{H2F1}, A_{H1F2}, and A_{F1F2}. The inference process can be modeled with the link distance (L) which indicates the minimum number of links between F1 and F2 in the social network. The shortest link distance, L equals to the sum of the number of links from F2 to H2 (between-group) and the number of links from H2 to F1 (within-group). Since we controlled the between-group distances as one, the brain areas encoding L also can be estimated by the GLM which includes D_{H2F1} as parametric regressors. In addition to the parametric regressors of distance, all GLMs were also included the rank of the task-relevant dimension and the rank of the task-

irrelevant dimension of presenting faces as additional regressors at the time of F1 and F2 presentation ($F1_R$, $F1_I$, $F2_R$, and $F2_I$). The onsets of the contextual dimension cue, F3 presentation, and button presses were also entered as additional regressors in all GLMs.

For the univariate neural model comparisons, we first estimated the log-likelihood of each of GLMs. Following previous work (Kumaran, Banino, Blundell, Hassabis, & Dayan, 2016; Niv et al., 2015; Wilson & Niv, 2015), the log-likelihood (LL) of each of the models was calculated (**Eq.2**) separately for the *a priori* anatomically defined ROIs – the EC and vmPFC/mOFC.

$$LL = n\left(\ln\sqrt{2\pi\sigma^2} + 0.5\right)$$

Eq. 2

where *n* is the total number of scans, and σ_2 is the variance of the residuals after subtracting the best-fit linear model. Considering that the linear model provides the maximum likelihood solution to each model with Gaussian-distributed noise, the likelihood was computed from residuals in the ROIs after subtracting the best-fit linear model. Since all models had the same number of parameters, their likelihoods could be directly compared to ask which model accounted best for the neural activation patterns. We further entered the LL to Bayesian model selection (BMS) to compare the goodness of the model with the exceedance probability (XP).

Regions of Interest (ROI) analyses

The ROIs were defined in the bilateral HC (Yushkevich et al., 2015), bilateral EC (Amunts et al., 2005; Zilles & Amunts, 2010), bilateral amygdala (AM) (Tzourio-Mazoyer et al., 2002), and bilateral vmPFC/mOFC (Neubert et al., 2015) using probabilistic map of anatomical ROIs. We also included additional ROIs in the bilateral primary motor cortex (M1) (Glasser et al., 2016) as control regions. Note that ROIs were independently defined from the current task. For the display purpose, all statistical parametric maps presented in the manuscript are unmasked.

ROI based representational similarity analysis (RSA)

Neural representation of social hierarchies while inferring relative social status between novel pair individuals

In hypothesis-driven analyses, we performed a representational similarity analysis (RSA) (Kriegeskorte, 2008; Nili et al., 2014) to test whether the *a priori* ROIs contain a reliable structure with respect to the social hierarchies. To test this hypothesis, we extracted the brain activity of each of the individuals from their unsmoothed beta maps. We estimated the patterns of brain

activity from each of ROIs. We averaged the neural activity patterns elicited while each of the individual faces has been shown at the time of F1 or F2. These neural representations were separately estimated according to which social hierarchy dimension was relevant to the current task. Reliability of data was improved by applying multivariate noise normalization (Walther et al., 2016). We quantified the representational similarity for the two blocks using the Mahalanobis distance between the activity patterns which generated a 24×24 representational dissimilarity matrix (RDM; 12 non-hub individuals were presented in each of two-dimension; **Fig. 4A**). These processes were repeated per ROI. We validated that the RDM estimated from the brain activity patterns in each of the ROIs discriminates different face stimuli with good sensitivity using the exemplar discriminability index (EDI) (Nili, Walther, Alink, & Kriegeskorte, 2016), which is defined as the average of the pattern dissimilarity estimates between different stimuli compared to the average of the pattern dissimilarity estimates between the same stimuli. We confirm that EDI in all ROIs was positive (one-sample t-test, $p<0.01$) suggesting that the different sets of stimuli were discriminable based on their multivariate activity patterns.

We predicted the RDM estimated from the patterns of neural activity in *a priori* ROIs by several model-based predictors (Model RDM; **Fig.3 B**). The model RDMs included (1) pairwise Euclidean distances between individuals in 2-D social space (*E*); (2) pairwise rank difference between individuals in the task-relevant hierarchy (*D*); (3) the context of which social hierarchy dimension the face stimulus was presented in (*C*, task-relevant dimension); (4) which was the group the stimulus belonged to during training (*G*).

In addition to the model RDMs, using partial correlation, we also tested for the effect specific to E while controlling the confounding covariance between E and D. Specifically, we estimated the extent to which the RDM estimated in each ROI was explained by E'. E' indicates the pairwise Euclidean distances between individuals (E) while regressing out its partial correlation with the pairwise rank difference in task-relevant hierarchy (D):$E' = E - DD^{+}E$ where D^{+} is the Moore-Penrose generalized matrix inverse ($D^{+} = pinv(D)$, **Supplementary Fig.8A**). This partial correlation method gives an advantage over the other methods such as orthogonalization which often loses the original structure (**Fig.S8A**). Note that, as Supplementary Fig.7A shows, E' differs from $\mathrm{E}^{\mathrm{Orth}}$ which indicates E orthogonalized by D using the Gram-Schmidt method or the pairwise rank difference in task-irrelevant hierarchy (I). After regressing out the partial correlation, the predictors are independent from each other while preserving high correlation with its original structure. That is, E' do not correlates with D while it still highly correlates with E (The right panel in **Fig.S8A**).

The extent to which the brain RDM of each ROI was explained by the model RDM was estimated with the rank correlation (Kendall's τ_A). This effect was further tested at the group-level with the Wilcox signed-rank test across participants. The ROI based analysis treats data from a whole ROI. Therefore, correction for multiple comparisons was made by the number of ROIs as well as by the number of comparisons. We reported the results corrected for family-wise error (FWE) for multiple comparisons with the Holm-Bonferroni method.

Last, we examined the relationship between pattern dissimilarities in each of the ROIs to pairwise Euclidean distances (E), the pairwise task-relevant rank differences (D**)**, and the pairwise task-irrelevant rank differences (I) of all between individuals in the social cognitive map. The brain RDM of each participant was normalized into a range between 0 and 1. The upper triangular part of the normalized 24×24 RDM was assorted by the distance measure in each model RDM which indicates the model prediction of representational dissimilarity. We estimated the mean pattern dissimilarity per bin across participants. This analysis was only performed for visualization purpose (**Fig.3E** for E; **Fig.S8C** for D; and **Fig.S8D** for I). We did not make any statistical interpretation based on this analysis.

Searchlight based RSA

The searchlight RSA was performed to examine the brain areas in which the activity patterns represent the structure of the social hierarchies outside of the HPC-ERC system. Moreover, the searchlight analysis allows us to examine to what extent the model RDM explains the neural representation dissimilarity with a fixed number of voxels examined across regions. We defined a sphere containing 100 voxels around each searchlight center voxel. Consistent with the ROI analysis, we estimated the neural activity patterns elicited while each of the individuals has been presented at the time of F1 or F2 from each of the searchlights. These neural representations were separately estimated according to which social hierarchy was relevant to the current task. The dissimilarity matrices were quantified with Euclidean distance between neural patterns estimated form different blocks. For each searchlight, therefore, a 24×24 dissimilarity matrix was generated based on neural activities elicited by each of face stimuli in two different task-relevant dimensions. We used the same predictors (i.e. model RDMs) that we used for ROI-based RSA analysis to estimate the neural representational dissimilarity across searchlights with Kendall's τ_A rank correlation. Given the need to assess specific neural dissimilarity representing E, we also used partial correlation, E' while controlling for its covariance with D. The computed Kendall's τ_A values were then mapped back on the central voxel, allowing continuous mapping of information in the whole-brain per subject. These images were further

smoothed using an 8-mm full-width at half maximum (FWHM) Gaussian kernel and Fisher's Z transformed. We further performed one-sample t-tests for a group-level analysis. We corrected for multiple comparisons using threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009) with 1000 times of simulation. We reported the results corrected for family-wise error (FWE) for multiple comparisons ($p_{TFCE}<0.05$).