Behavioral/Cognitive

Distance and Direction Codes Underlie Navigation of a Novel Semantic Space in the Human Brain

Simone Viganò and Manuela Piazza
Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto (Tn), Italy

A recent proposal posits that humans might use the same neuronal machinery to support the representation of both spatial and nonspatial information, organizing concepts and memories using spatial codes. This view predicts that the same neuronal coding schemes characterizing navigation in the physical space (tuned to distance and direction) should underlie navigation of abstract semantic spaces, even if they are categorical and labeled by symbols. We constructed an artificial semantic environment by parsing a bidimensional audiovisual object space into four labeled categories. Before and after a nonspatial symbolic categorization training, 25 adults (15 females) were presented with pseudorandom sequences of objects and words during a functional MRI session. We reasoned that subsequent presentations of stimuli (either objects or words) referring to different categories implied implicit movements in the novel semantic space, and that such movements subtended specific distances and directions. Using whole-brain fMRI adaptation and searchlight model-based representational similarity analysis, we found evidence of both distance-based and direction-based responses in brain regions typically involved in spatial navigation: the medial prefrontal cortex and the right entorhinal cortex (EHC). After training, both regions encoded the distances between concepts, making it possible to recover a faithful bidimensional representation of the semantic space directly from their multivariate activity patterns, whereas the right EHC also exhibited a periodic modulation as a function of traveled direction. Our results indicate that the brain regions and coding schemes supporting relations and movements between spatial locations in mammals are "recycled" in humans to represent a bidimensional multisensory conceptual space during a symbolic categorization task.

Key words: **Concepts; entorhinal cortex; navigation; Semantic; ventro-medial prefrontal cortex**

Significance Statement

The hippocampal formation and the medial prefrontal cortex of mammals represent the surrounding physical space by encoding distances and directions between locations. Recent works suggested that humans use the same neural machinery to organize their memories as points of an internal map of experiences. We asked whether the same brain regions and neural codes supporting spatial navigation are recruited when humans use language to organize their knowledge of the world in categorical semantic representations. Using fMRI, we show that the medial prefrontal cortex and the entorhinal portion of the hippocampal formation represent the distances and the movement directions between concepts of a novel audiovisual semantic space, and that it was possible to reconstruct, from neural data, their relationships in memory.

Introduction

In humans, the ability to represent and navigate the surrounding physical space is supported by the activity of the hippocampal formation and of the medial prefrontal cortex (mPFC), where place cells (O'Keefe and Dostrovksy, 1971) and grid cells (Hafting et al., 2005) fire to represent the current location of an individual (Ekstrom et al., 2003; Jacobs et al., 2013). Grid cell activity, in particular, can be inferred using BOLD fMRI when participants move in virtual reality environments (Doeller et al., 2010). A recent proposal posits that our species uses the same neuronal machinery to support internal representations ("cognitive maps") of nonspatial memories and experiences (Tolman, 1948; Stalnaker et al., 2015; Behrens et al., 2018; Bellmund et al., 2018). This proposal comes from a set of complementary observations. First, from the neuroanatomical point of view, the same brain regions where spatially tuned neurons have been recorded [hippocampus, entorhinal cortex (EHC), and medial prefrontal cor-

Received July 31, 2019; revised Dec. 17, 2019; accepted Jan. 15, 2020.

Author contributions: S.V. and M.P. designed research; S.V. performed research; S.V. analyzed data; S.V. and M.P. wrote the paper.

The study was supported by the Italian Minister for Education, Universities, and Research (MIUR) project "Dipartimenti di eccellenza." We thank Marco Buiatti for fruitful discussions, Maria Ravera for support during behavioral training, and the anonymous reviewers who shared crucial comments on previous versions of the manuscript.

The authors declare no competing financial interests.

Correspondence should be addressed to Simone Viganò at simone.vigano@unitn.it.

https://doi.org/10.1523/JNEUROSCI.1849-19.2020

tex] also encode nonspatial relations among stimuli (temporal, associative, hierarchical) and are recruited during abstract reasoning and decision-making (Schuck et al., 2016; Kaplan and Friston, 2019). Second, from a neurofunctional point of view, during tasks that bear little, if any, similarity with physical navigation, such as imagined navigation (Bellmund et al., 2016), visual search (Julian et al., 2018; Nau et al., 2018), or processing of morphing visual objects (Constantinescu et al., 2016) and odors (Bao et al., 2019), the activity of these regions is modulated by the same sixfold periodicity that is characteristic of grid cells observed during spatial navigation.

Because the hallmark of the human species is the use of language to organize conceptual representations, we asked whether we can find traces of space-like coding of semantic knowledge in the human brain. In humans, concepts are representations that typically refer to *categories* of objects or events (the category, "cat") and that can be equally activated when processing individual instances of the object/events themselves (a specific picture of a specific cat) or their corresponding arbitrary *symbol* (the written or spoken word/cat; Quiroga et al., 2005; Fairhall and Caramazza, 2013). In cognitive (neuro)science, scholars tend to interpret concepts (represented by words) as regions or points in an internal space (the "semantic space"), with proximities reflecting similarities in meaning (Gärdenfors, 2000; Borghesani and Piazza, 2017). If the human brain uses the same neuronal machinery to represent and navigate both the physical and the semantic space, two predictions follow. First, the brain regions previously shown to be involved in mapping the physical space (the hippocampal formation and the medial prefrontal cortex in particular) should also hold a map of the semantic space. Second, in those regions, the neural coding schemes characterizing spatial navigation (distance-based and direction-based codes) should also support navigation in semantic space. To date, the evidence supporting these two predictions is still missing, mostly because it is extremely difficult to reduce complex human semantic representations to low-dimensional spaces that allow a straightforward comparison with the 2D or 3D navigable physical environment.

To overcome this problem, we created a novel, artificial, highly controlled semantic space composed of audiovisual objects parsed into four categories by means of linguistic labels (words; Fig. 1*A*,*B*; see Materials and Methods). Participants learned to assign each of the objects (resulting in a specific combination of visual and audio features) to a particular labeled category during 9 d of behavioral training. Before and after training, they were presented, during an fMRI scanning session, with pseudorandom sequences of objects and words while performing a one-back task (see Materials and Methods). We reasoned that the activity evoked by stimuli referring to different regions of the semantic space should reflect the *distance* existing between them: the closer two concepts are in the semantic space, the closer (or similar) their representations should be. Additionally, we reasoned that subsequent presentations of words and objects referring to different categories implied a specific *direction* traveled within the semantic space (Fig. 1*C*; see Materials and Methods). Thus, this novel semantic space allowed for testing the presence of both distance-based and direction-based neural codes.

Materials and Methods

Participants. The study included 25 right-handed adult volunteers (15 females and 10 males; mean age = 22.20 years, SD = 2.74). All participants gave written informed consent, underwent screening to exclude incompatibilities with the MRI scanner, and were reimbursed for their time. The study was approved by the ethics committee of the University of Trento (Italy).

Semantic space. We developed a set of 16 novel multisensory objects and orthogonally manipulated the size of an abstract shape (Fig. 1*A*,*B*) and the pitch of a sound the objects produced during a small animation. A total of four size and pitch levels were used for each participant, leading to a stimulus space where each object represented the unique combination of one size and one pitch level. The values of these two features were selected for each participant on the first day of the experiment, following a brief psychophysical validation consisting of a QUEST adaptive staircase method (Watson and Pelli, 1983). Using a two-stimuli comparison task for each sensory modality, we calculated subject-specific sensitivity as the minimum appreciable increment [just noticeable difference (JND)] from a reference value (size: visual angle of 5.73°, pitch: frequency of 800 Hz) leading to 80% of correct responses. For each sensory modality, four subject-specific feature levels were calculated, applying the logarithmic Weber–Fechner law and selecting values at every three JNDs, to ensure that feature levels were equally distant in their representational space. Objects were animated by simulating a small squeezing; their presentation lasted a total of 750 ms, and sounds were presented at the apex of the squeezing period. The object space was divided into four categories based on the combination of two sensory boundaries (Fig. 1*B*). The categorical membership of each object, as well as objects' unique multisensory identities, could thus be recovered only by integrating the two sensory features. We assigned a novel word to each category (Fig. 1*C*): KER (small size and low pitch), MOS (big size and low pitch), DUN (small size and high pitch), GAL (big size and high pitch).

Stimuli presentation. Stimuli were presented foveally using MATLAB Psychtoolbox (MathWorks) in all experimental phases, at a distance of ~130 cm. Multisensory objects subtended a different visual angle for each size level and a different frequency for each pitch level, ranging from an average of 5.73° and 800 Hz for level 1 (size and pitch, respectively) to an average of 8.97° and 973.43 Hz for level 4. Each word subtended a visual angle of 3.58° horizontally and 2.15° vertically and was presented with black Helvetica font on a gray background.

Experimental sessions. The experiment consisted of three parts: prelearning fMRI, behavioral training, and postlearning fMRI. During prelearning fMRI, participants were exposed for the first time to the new multisensory objects and to the abstract names. This allowed recording of the patterns of neural activity evoked by the stimuli when they did not share any relationship. Starting with the following day, subjects underwent nine sessions of behavioral training outside the scanner. The aim was to teach them the object–name correspondence, an operation requiring parsing of the object space into four categories and connecting each symbol (word) to its meaning (the correct category exemplars). Finally, during the postlearning fMRI, they were again exposed to the same objects and words, now probing their mutual correspondence and allowing us to record the updated cortical activity. On average, the second fMRI session occurred 9.86 d (SD = 1.4 d) after the first one. All tasks are described here. During both fMRI sessions, and during the first eight training days, we used 8 of 16 objects available in each subject's stimulus space (objects: 1-3-6-8-9-11-14-16); the remaining eight were used only during the ninth training session to test for generalization (see Results, Behavioural results).

fMRI tasks. During the first fMRI session, participants were presented with the multisensory objects and the four abstract words in pseudorandom order and performed a one-back task on stimulus identity. They were instructed to press a button when they detected a rare immediate repetition of the very same stimulus (either multisensory object or word). Each stimulus was presented for either 750 ms (objects) or 500 ms (words), with a variable inter-stimulus interval of 4 ± 1.5 s, during which a blue fixation cross was presented. There were four runs, each one lasting ~7 min. Within a run, each stimulus (eight objects and four words) was repeated six times, resulting in 72 trials per run. There was one target event (one-back repetition) per stimulus for a total of 12 of 72 (~17%) expected responses per run. During the second fMRI session, participants were again presented with the same multisensory objects and the four abstract words in pseudorandom order, as in the pretraining fMRI session, but now they performed a one-back task on word– object corre-

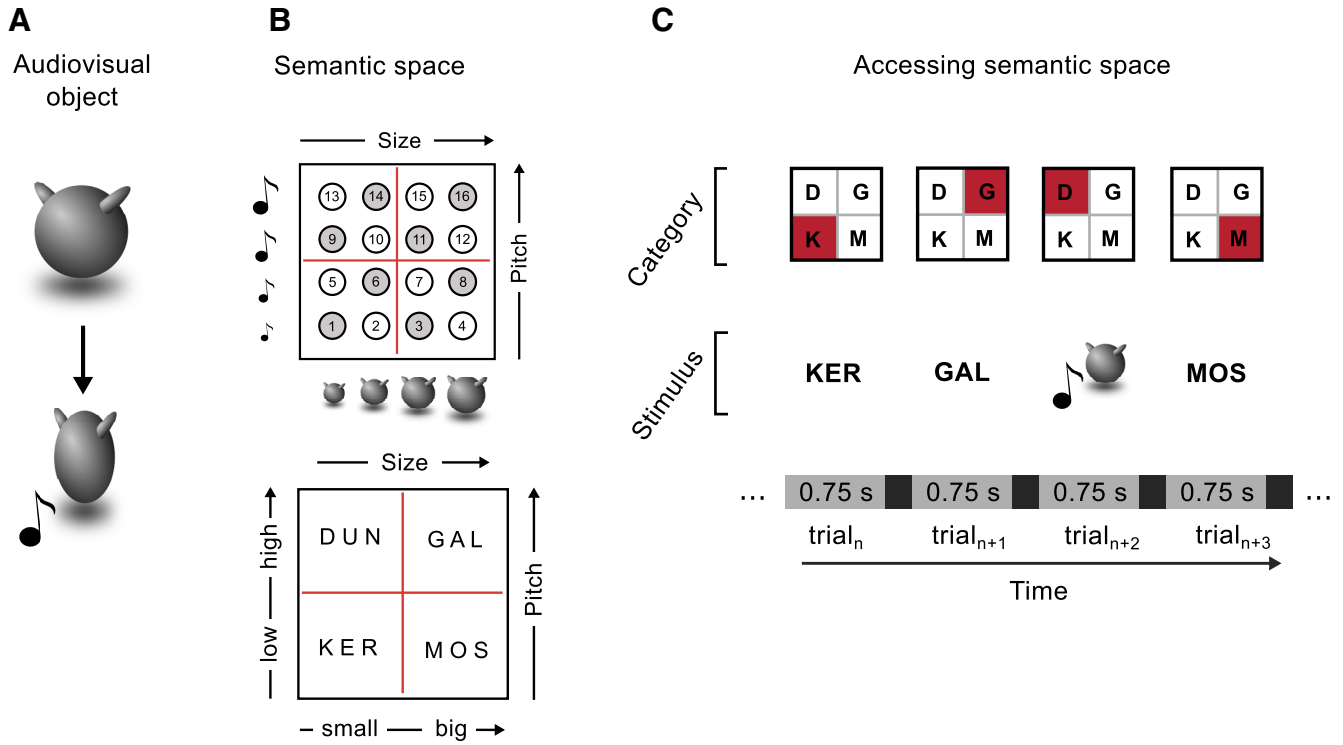

Figure 1. Methods. ***A***, Exemplar of audiovisual object. ***B***, Sixteen multisensory objects are divided into four categories by means of abstract words: this creates a novel multisensory semantic space. Objects in gray are the ones used during the fMRI sessions (see Materials and Methods) ***C***, Subsequent presentations of either objects or words imply a movement between the regions of the semantic space. These movements are characterized by a specific distance and direction.

spondence, where they had to correctly associate each object with the corresponding name. This task could not be performed before learning given the absence of any categorical knowledge for our stimulus space. Participants were instructed to press the button any time a multisensory object was followed by the corresponding name (e.g., object 1 followed by the word KER) and vice versa (e.g., word KER followed by object 1), thus requiring access to a newly learned symbolic identity. This resulted in a total of 16 target events (~22%) per run. The number of runs, trials, and stimuli repetitions matched the one-back task on stimulus identity on the first fMRI day. This task guaranteed that subjects accessed the categorical representations of words and objects, without suggesting any spatial arrangement.

Behavioral training. Participants underwent nine daily sessions of behavioral training, aimed at making them learn the correct name of each object. Each behavioral session was ~10 min long and was divided into four miniblocks of 20 trials each, for a total of 80 trials. It started with a brief presentation of the objects as exemplars of the four categories (KER, MOS, DUN, GAL). After this familiarization phase, each trial consisted of an object presentation (750 ms) followed by a fixation cross (500 ms) and by the presentation of the four possible names in random order, from left to right. Each object was presented 10 times per training session. Participants were instructed to press one key on the keyboard to select the correct name. They were asked to respond as fast as possible, but no time limits were imposed. After their response, immediate feedback appeared on the screen for 1000 ms, indicating the accuracy of the choice with the words "Correct!" or "Wrong!" In the case of a wrong answer, the feedback also showed the correct object name to speed up the learning process. After each miniblock, participants were provided with the cumulative percentage accuracy. Starting from the seventh training session, the trial-by-trial feedback was removed, and participants could rely only on the block-by-block cumulative feedback. For the first 8 d of training, participants were presented with the same eight objects used in the two fMRI sessions. On the last training day, without being notified of the change, they were presented with all 16 objects. This allowed us to test for generalization of the categorical rule to new exemplars (here represented by objects 2-4-5-7-10-12-13-15), which would be a key ingredient of an efficient semantic representation. For this last session, the miniblock's number was kept at four but the number of trials was doubled, resulting in a total testing time of ~20 min.

Neuroimaging acquisition and preprocessing. Data were collected on a 4 T Bruker scanner (Bruker BioSpin) with standard head coil at the Center for Mind/Brain Sciences, University of Trento, Italy. Functional images were acquired using EPI T2*-weighted scans. Acquisition parameters were as follows: TR = 3 s; TE = 21 ms; flip angle = 81°; FOV = 100 mm; matrix size = 64 × 64; number of slices per volume = 51, acquired in interleaved ascending order; voxel size = 3 × 3 × 2 mm. T1-weighted anatomical images were acquired twice per participant (prelearning and postlearning) with an MP-RAGE sequence, with 1 × 1 × 1 mm resolution. Functional images were preprocessed using the Statistical Parametric Toolbox (SPM8) in MATLAB. Preprocessing included the following steps: realignment of each scan to the first of each run, coregistration of functional and session-specific anatomical images, segmentation, AND normalization to the Minnesota National Institute (MNI) space. No smoothing was applied.

Distance-based adaptation analysis. First of all, we assessed what brain regions, after learning, represented the reciprocal distances between the four concepts. We did that by means of fMRI adaptation, reasoning that a large distance (e.g., from KER to GAL) traveled in the conceptual space should result in a higher release from adaptation compared with a small distance (e.g., from KER to MOS) due to the higher number of sensory features differentiating between the two concepts. Previous studies (Henson et al., 2002; Henson, 2016) highlighted that adaptation effects are more easily detectable when the repetition is not task relevant; therefore, we focused on transitions between two stimuli that belonged to two different categories when participants did not have to press the response button key. Functional images for each participant individually were analyzed using a general linear model (GLM). For each run, 14 regressors were included: one regressor for each pair of trials of no interest where no movement happened (e.g., two subsequent stimuli referring to the same category); two regressors of interest modeling pairs of trials where either a small or a large movement occurred in the conceptual space (Fig. 2*A*); one regressor indicating that there was a change in the presentation

modality (from object to word or vice versa); one regressor for motor response; six regressors for head movements (estimated during motion correction in the preprocessing); three regressors of no interest (constant, linear, and quadratic). Baseline periods were modeled implicitly, and regressors were convolved with the standard HRF (without derivatives). A high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. We applied group-level analysis within SPM to find brain regions showing a significant adaptation effect for trials in which the movement covered a large distance over those in which the movement covered a small distance [familywise error (FWE) correction for multiple comparisons at cluster level was applied at $\alpha = 0.05$].

Distance-based representational similarity analysis. Next, we applied a distance-dependent model-based multivariate pattern analysis (MVPa), which is complementary to the distance-dependent adaptation analysis and could give complementary evidence for a distance-based representation. To do so, we ran a second GLM. For each run, 22 regressors were included: one regressor for each of the eight multisensory objects (resulting in eight regressors); one regressor for each of the four words (resulting in four regressors); one regressor for motor response; six regressors for head movements (estimated during motion correction in the preprocessing); three regressors of no interest (constant, linear, and quadratic). Again, baseline periods were modeled implicitly, regressors were convolved with the standard HRF without derivatives, and a high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. We thus obtained one β map for each stimulus (object or word) and run. We used these β maps to conduct a model-based representational similarity analysis (RSA; Kriegeskorte et al., 2008) within a whole-brain searchlight (radius of the sphere = 3 voxels). We averaged the β maps for all the stimuli that belonged to the same concept (e.g., objects that are a KER and the word "KER") and we extracted, from each sphere of the searchlight, the neural dissimilarity matrix (DSM; 1-Pearson's correlation) to reveal their distances in the multivariate neural representational space. A sphere was centered in every voxel of the subject-specific and session-specific datasets, following previous searchlight studies (Connolly et al., 2012). Within each sphere, we correlated the Fisher transformed DSM to the predicted matrix representing the distances between concepts (Fig. 2*D*). We used SPM to test for group-level effects after subtracting the results of two additional searchlights (with matching parameters) that used model-based RSA to look for brain regions responding to differences in either size or pitch of the multisensory objects: this was a necessary step to exclude that the multivariate correlation score we obtained could be explained by a low-level perceptual coding of differences between objects. FWE correction was applied at $\alpha = 0.05$ to correct for multiple comparisons at the cluster level.

Direction-based RSA. Next, we tested the presence of a direction-based neuronal code, asking whether BOLD activity evoked during the transition between two stimuli referring to different concepts was modulated by the direction of the movement in the semantic space. To do that, we ran a third GLM, this time modeling the directions of movement between concepts instead of the distance between them. For each run, 20 regressors were included: eight regressors corresponding to the eight possible directions of movement within the conceptual space, arbitrarily referenced to the horizontal axis; one regressor modeling subsequent presentation of two stimuli that belonged to the same conceptual region, corresponding to no movement across the conceptual environment; one regressor for changes in presentation modality (e.g., from object to word or vice versa); one regressor for participants' response; six regressor for head movements (estimated during motion correction in the preprocessing); three regressors of no interest (constant, linear, and quadratic). Baseline periods were modeled implicitly, and regressors were convolved with the standard HRF without derivatives. A high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. We thus obtained one β map for each movement direction for each run. In the ROI defined by previous analysis, we applied an extension of the similarity-based multivariate approach of Bellmund et al. (2016) to test for the existence of a hexadirectional code in our data, most likely originating from the activity of grid cells (Doeller et al., 2010). This approach, contrary to the univariate quadrature filter procedure (Doeller et al., 2010; Constantinescu et al., 2016), is not based upon the estimation of the preferred grid orientation of each voxel, but solely assumes that it varies across voxels (Fig. 3*A*). This assumption is derived from empirical data both using fMRI in humans (Doeller et al., 2010; Nau et al., 2018) and using electrophysiological recording in rodents, indicating that the grid orientation varies in a step-like fashion across different portions of the entorhinal cortex (Stensola et al., 2012). Starting from this assumption, we reasoned that two movement directions, φ and φ', in the interval 0°–359° can be expressed as more or less similar in an n-fold periodic space by calculating $\mathrm{mod}(\varphi\text{-}\varphi', \theta)$, where θ indicates the angle of the periodic (n-fold) grid for which we want to test the modulation. In the case of a grid-like signal, corresponding to a sixfold periodicity, $\theta = 60°$; therefore, two directions perfectly aligned with a periodicity of 60° would have $\mathrm{mod}(\varphi\text{-}\varphi', 60°) = 0$. However, if the two directions are not perfectly aligned in the n-fold symmetry, the result of the mod() function indicates the angular distance from perfect alignment, which is proportional to their *dis*similarity. Note that the logic used here is an extension of the multivariate approach used in typical RSA analyses (Haxby et al., 2001; Kriegeskorte et al,. 2008) with the supplemental assumption of a specific response tuning function of the underlying neuronal population. We thus computed all 8×8 pairwise comparisons between our sampled movement directions to obtain a model of their predicted sixfold dissimilarity, corresponding to the angular deviation in the 60° periodic space. Next, we applied model-based RSA correlating the sixfold model to the Fisher transformed neural DSM constructed by computing similarity distance (1-Pearson's r) between any pair of distributed activity patterns in the ROI (Fig. 3*A*). We computed the correlation between neural data and the model using Pearson's r. To investigate whether this modulation was detectable at the whole-brain level, we used the CoSMoMVPa toolbox (Oosterhof et al., 2016) to implement this analysis in a whole-brain searchlight (radius = 3 voxels) to find cortical regions responding to the sixfold rotational symmetry. A sphere was centered in every voxel of the subject-specific and session-specific datasets, following previous searchlight studies (Connolly et al., 2012). Within each sphere, we conducted the model-based grid-RSA, storing the resulting correlation score in the center voxel, as a summary of the information present in the surrounding sphere. Motivated by previous findings (Doeller et al., 2010) indicating that the right entorhinal cortex shows sixfold modulation during spatial navigation, we applied a small volume correction at the group level using a spherical ROI of radius three voxels centered at MNI coordinates 30, 3, −32, derived from the study by Doeller et al. (2010).

Multidimensional scaling. Finally, and mainly for visualization purposes, we used the neural DSM extracted from the regions obtained during the distance-dependent and the direction-dependent analyses and averaged across participants to reconstruct, using multidimensional scaling as implemented in MATLAB (functions cmdscale and voronoi, for Voronoi tessellation), the most faithful underlying bidimensional representation of the semantic space, to visualize the spatial arrangement of the four concepts starting from real neural data, for both the prelearning and postlearning datasets (see Fig. 4).

Data and code availability. The data collected and the code used are available from the corresponding author upon reasonable request.

Results

The main goal of the study was to investigate whether and where a map of a novel, two-dimensional conceptual space is represented in the brain of healthy adult participants. The key ingredients of such a map are those of reflecting the distances and directions existing between the locations (the concepts) of the space it represents. To test this, we used a combination of univariate (adaptation) and multivariate (RSA) techniques (see Materials and Methods).

Behavioral results

During the learning phase outside the scanner, participants were trained for eight daily sessions to parse a multisensory objects space into categories using words, by means of a delayed match-to-category-name task on eight specific multisensory objects (see

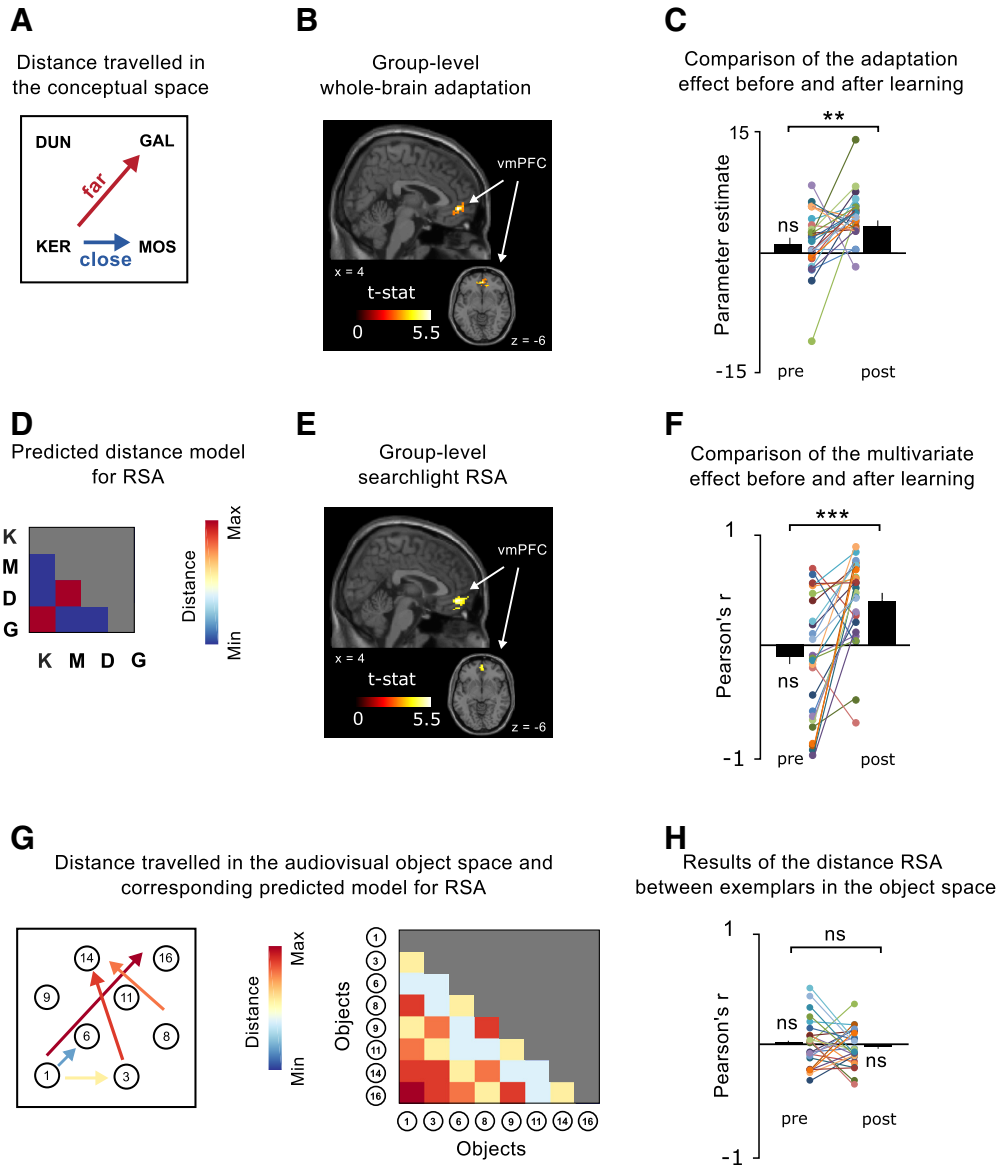

Figure 2. Results of the distance analysis. ***A***, Moving in the semantic space implies covering different distances. ***B***, Results of a whole-brain adaptation reveal a significant cluster in the mPFC reflecting distances between semantic regions. ***C***, This effect was not present before learning. ***D***, We additionally tested the same effect with a complementary multivariate approach: representational similarity analysis. The more two regions are distant in the semantic space, the more different the activity they evoke. ***E***, Results of a whole-brain RSA searchlight reveal a significant cluster in the mPFC, consistent with the adaptation analysis. ***F***, The multivariate effect was not present before learning. ***G***, Model of predicted distances in the audiovisual object space. ***H***, The model of distances between audiovisual object exemplars does not correlate with activity in the mPFC, neither before nor after learning. $^{**}p < 0.005$, $^{***}p < 0.001$. ns, not significant.

Materials and Methods). During the last training session, and without being notified, they were also presented with eight novel stimuli that they never saw before. These consisted of specific combinations of size and pitch that were absent in the training set, and they were introduced to verify the emergence of a real categorical rule-based representation of the semantic space, and not of a mere learning of the mapping between individual exemplars and the corresponding names. The learning trajectory indicated a significant increment in performance from session 1 to session 8 (session 1: 60 ± 18%; session 8: 89 ± 8%; paired *t* test: $t_{(24)} = 8.58$, $p = 8.86 \times 10^{-9}$). Performance collected on session 9 confirmed the successful learning and generalization of the categories (performance training set: 87 ± 7%; different from chance, $t_{(24)} = 40.29$, $p = 1.48 \times 10^{-23}$; performance generalization set: 73 ± 11%, difference from chance, $t_{(24)} = 21.49$, $p = 3.45 \times 10^{-17}$).

After learning, participants underwent an fMRI session performing a one-back-category-name task (see Materials and Methods). Performance in the scanner was high (hit = 84 ± 10%, correct rejection = 97 ± 1%). At the end of the experiment, no participant reported having explicitly conceptualized the stimulus space using any kind of spatial arrangement.

Neuroimaging results

A distance-dependent code

Evidence from adaptation. First, we investigated whether and where in the brain the newly acquired semantic information was represented using a distance-dependent code. We reasoned that a subsequent presentation of two stimuli belonging to two different categories would cause an adaptation of the BOLD signal that would be proportional to the distance between the two concepts in the two-dimensional concept space [see Piazza et al. (2004) for

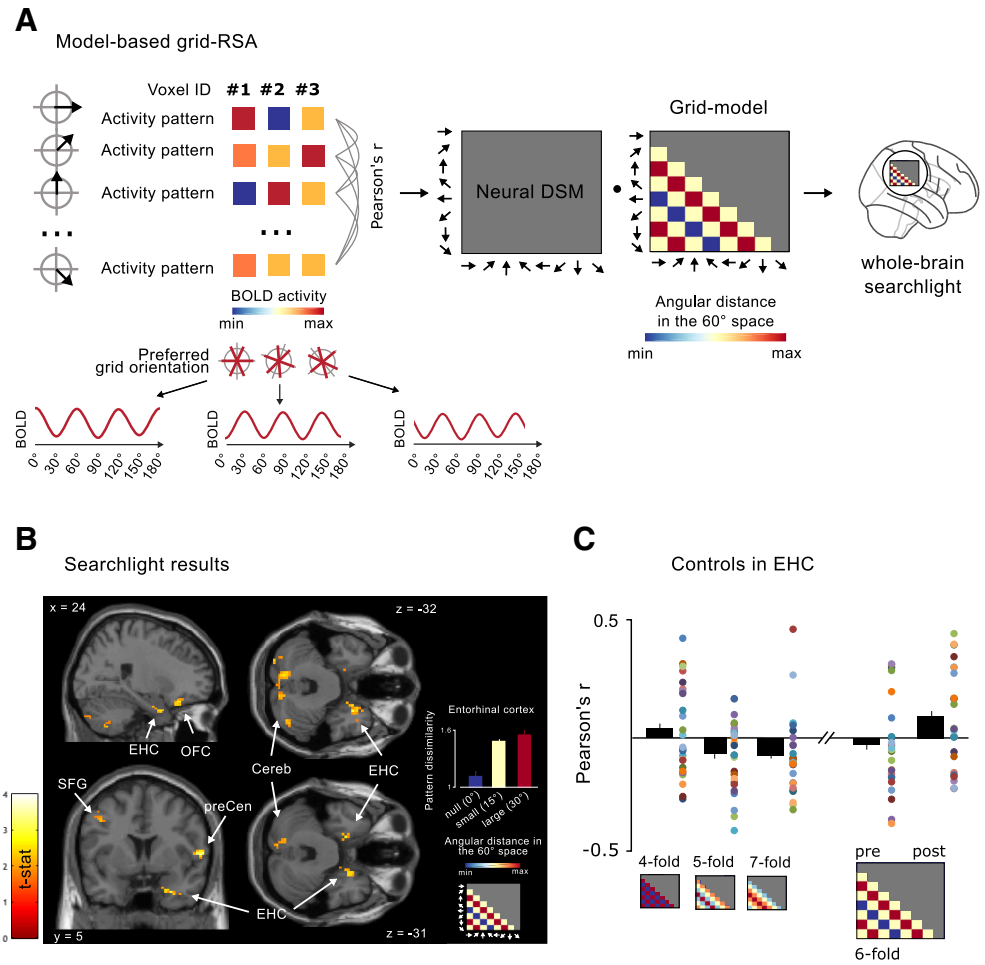

Figure 3. Methods and results of the model-based grid-RSA analysis. ***A***, Left, Different movement directions (represented as black arrows in gray circles) elicit a different multivariate activity pattern in a brain region, where each voxel shows higher or lower BOLD signal as a function of its preferred grid orientation. Here, we represent three putative voxels with different preferred grid orientations. The model-based grid-RSA method capitalizes on this variability and produces a predicted dissimilarity matrix ("grid-model") representing the predicted pairwise correlations between the activity patterns evoked by the different directions as a function of their angular distance in the 60° periodic space. This sixfold grid-model is then correlated with the fMRI activity patterns evoked by the different movement directions to reveal the presence of a grid-like response. ***B***, Results of the whole-brain searchlight using the model-based grid-RSA with the sixfold periodic model. Threshold at $p < 0.05$ uncorrected for visualization. For illustrative purposes, we also show that the average dissimilarity between activity patterns evoked by different directions in the entorhinal cortex varies as a function of their angular distance modulo 60 (null = 0°, small = 15°, large = 30°). ***C***, Correlation with competing periodicities at fourfold, fivefold, and sevenfold. Sixfold modulation is absent before learning.

a similar parametric approach in one-dimensional space]. In our stimulus space, concepts were separated by two possible levels of distance: small (e.g., KER preceded by MOS) or large (e.g., KER preceded by GAL), corresponding to differences along one sensory dimension only or along both. Regions where the BOLD signal is affected by this difference should be detectable using fMRI adaptation, where the response to a given stimulus should be lower when that stimulus is preceded by another one with a small conceptual distance compared with a large distance. This analysis, applied at whole-brain level on postlearning fMRI data, revealed a single significant cluster in the mPFC ($MNI_{x,y,z} = 3, 47, -4$; $T = 6.73$; FWE corr.; Fig. 2*B*). To address the possibility that these results can be explained by the different number of stimulus properties being changed across conditions, we applied the same whole-brain adaptation analysis using prelearning data, where the stimuli presented to subjects were exactly the same, but the abstract mapping had not yet been learned. No significant cluster was found. To further explore this effect, we adopted an ROI approach restricting the adaptation analysis on the prelearning data to the BOLD signal in the medial prefrontal cluster, showing the distance effect postlearning. The mean parameter estimate across subjects extracted from the mPFC on prelearning was parameter estimate = 1.32 (SEM = 0.77, $t_{(24)} = 1.75$, CI = [−0.23, 2.88], $p = 0.09$), with a significant difference from postlearning (postlearning ≠ prelearning; $t_{(24)} = 2.94$, CI = [0.87, 4.98], $p = 0.007$; Fig. 2*C*).

Evidence from RSA. To confirm the previous result with an independent and complementary measure, we extracted the distributed activity patterns for each stimulus after running a second GLM (see Materials and Methods). Next, we implemented a whole-brain searchlight after excluding brain regions responding to differences along either size or pitch between the audiovisual objects (see Materials and Methods): in each sphere of the searchlight, we applied model-based RSA (Kriegeskorte et al., 2008) by correlating the Fisher transformed neural DSM (1-Pearson's *r*) extracted from the sphere to a model of the predicted distances in the conceptual space. We found two significant clusters: one in the mPFC ($MNI_{x,y,z} = 6, 50, -10$; $T = 5.71$; FWE corr.; Fig. 2*E*) and one in the precentral gyrus ($MNI_{x,y,z} = 57, 5, 24$; $T = 6.55$; FWE corr.). To address the possibility that these results could be explained by the amount of perceptual difference across stimuli, we applied the same RSA approach to prelearning data. With a

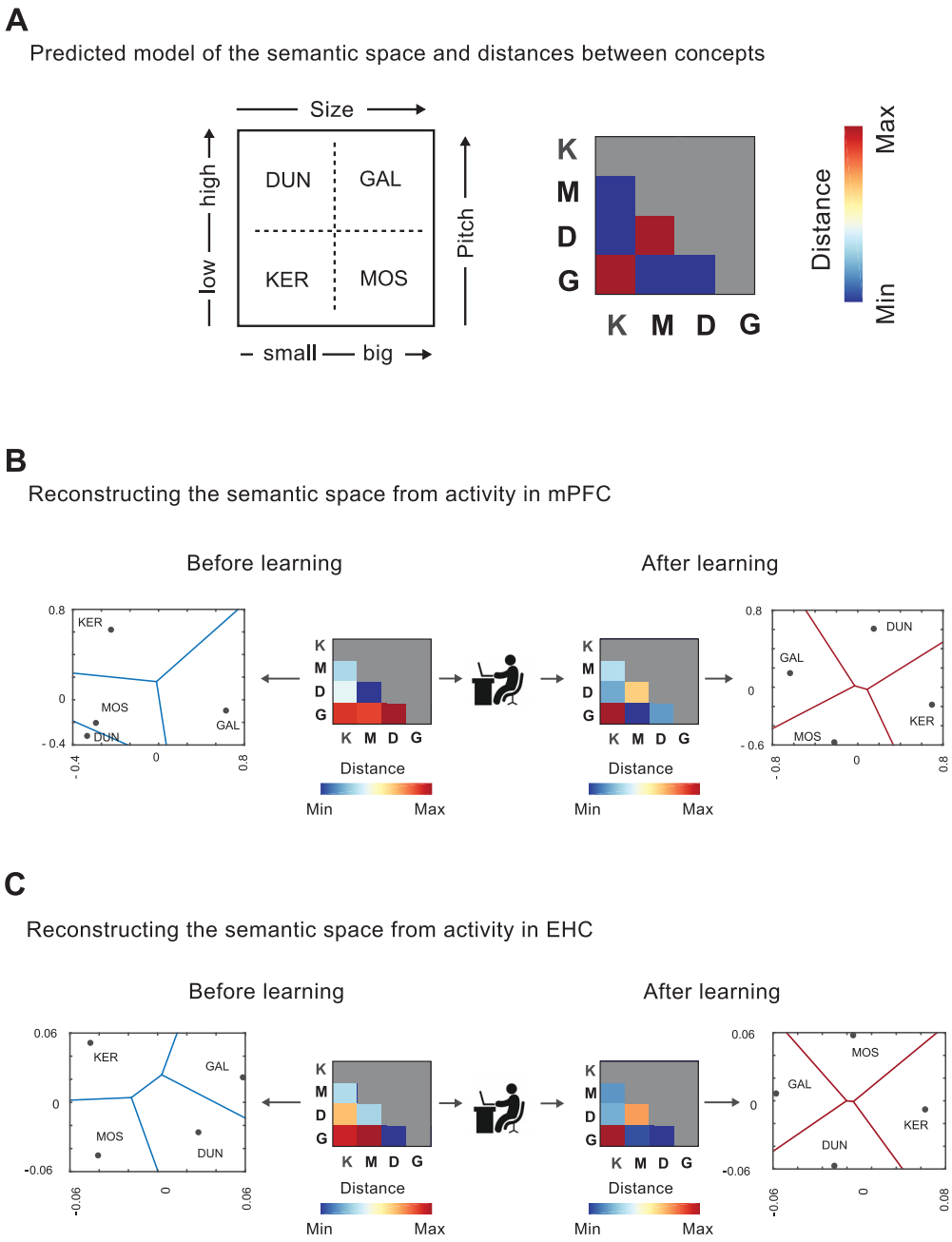

Figure 4. Reconstruction of the underlying spatial arrangement using multidimensional scaling. ***A***, Predicted model of the semantic space and relative distances between the four audiovisual concepts. ***B***, Recovery of the bidimensional representation from medial prefrontal distributed activity before and after learning. ***C***, Recovery of the bidimensional representation from right entorhinal distributed activity before and after learning. ***B***, ***C***, Blue and red lines are the results of Voronoi tessellation.

whole-brain searchlight analysis, no cluster was found significant. Using an ROI approach, the mean correlation of the BOLD signal across the mPFC voxels with the predicted model across subjects was not significant ($r = -0.10$; SEM = 0.06, $t_{(24)} = -0.9$, CI = [−0.33, 0.13], $p = 0.37$), and was significantly different from the post-training data (post-training ≠ pretraining; $t_{(24)} = 4.20$, CI = [0.26, 0.75], $p = 3.16 \times 10^{-4}$; Fig. 2*F*).

The distance-dependent code is specific to the conceptual space, and not to the audiovisual object space

Another possibility to explain our results would be that the mPFC reflects the sensory distances between object exemplars, which can also be conceived as points in a 2D audiovisual object space. We thus repeated the RSA, restricting it to individual objects (thus excluding word trials and not merging objects in classes), to test whether the multivariate activity in the mPFC significantly correlated to the model of the graded distances between objects (Fig. 2*G*). The results indicated no exemplar-based graded effect in postlearning ($r = -0.02$, SEM = 0.03, $t_{(24)} = -0.78$, CI = [−0.09, 0.04], $p = 0.44$) or in prelearning ($r = 0.008$, SEM = 0.05, $t_{(24)} = 0.17$, CI = [−0.08, 0.10], $p = 0.87$) and no significant difference between the two ($t_{(24)} = 0.65$, CI = [−0.07, 0.14], $p = 0.52$; Fig. 2*H*).

A direction-dependent code: evidence from RSA

The second key property of a movement in space is its direction. During navigation in physical space, it is thought that the activity of grid cells, whose tuning functions peak at multiple locations

following a sixfold periodicity, is key in coding movement direction. In a previous study (Constantinescu et al., 2016), it was observed that the BOLD signal in the mPFC as well as in other regions, including the entorhinal cortex, was modulated following a sixfold periodicity, typical of grid cells (Hafting et al., 2005; Doeller et al., 2010), when human participants processed morphing bird shapes varying continuously in their neck:legs ratio, therefore mimicking a movement in an artificial bidimensional "bird space" akin to real-world physical space. Although we could only sample eight movement directions with our design, we asked whether we could observe a similar modulation between our four discrete concepts that met both definitional criteria of human-like, high-level conceptual representations: that of being categorical, and that of having a meaning that is conveyed with a linguistic symbol, such as a word. We fit a new GLM to extract a β series for each movement direction across the four regions of the novel semantic space (see Materials and Methods), resulting in eight sampled directions. By building on findings that a hexadirectional code can be observed in multivariate activity patterns under highly controlled circumstances (Bellmund et al., 2016), we combined grid analysis with model-based RSA (Kriegeskorte et al., 2008). We computed all dissimilarity measures between directions, thus obtaining a model (matrix) of the relative distances to the hypothetical hexadirectional grid (see Materials and Methods; Fig. 3*A*). First, we used an ROI approach and used model-based RSA to test whether or not the neural dissimilarity matrix extracted from the mPFC fit with the model by computing Pearson's *r*. This was not the case ($t_{(24)} = 0.41$, $p = 0.68$). However, when we implemented the same analysis in a whole-brain searchlight, we found a significant cluster in the right entorhinal cortex [$MNI_{x,y,z} = 30, 5, -32$; $p < 0.005$; small volume corrected at $p < 0.05$ using an entorhinal ROI (see Materials and Methods)]. Additional clusters were observed in the left orbitofrontal cortex ($MNI_{x,y,z} = -15, 44, -20$), the left superior frontal gyrus ($MNI_{x,y,z} = -30, 23, 60$), and the right precentral gyrus ($MNI_{x,y,z} = 60, 5, 6$; all $p < 0.005$ uncorrected; Fig. 3*B*). Given previous studies showing directional modulation of BOLD signal as a function of movement direction in the entorhinal cortex during both spatial (Doeller et al., 2010) and nonspatial tasks (Bellmund et al., 2016; Constantinescu et al., 2016; Julian et al., 2018; Nau et al., 2018), and given the high proximity of our entorhinal peak ($MNI_{x,y,z} = 30, 5, -32$) to the one reported by Doeller et al. (2010) during spatial navigation ($MNI_{x,y,z} = 30, 3, -30$), we focused on this region for subsequent analyses. First of all, we verified that other periodicities (fourfold, fivefold, and sevenfold) did not account for the signal in this region (fourfold: $t_{(24)} = 0.35$, $p = 0.94$; fivefold: $t_{(24)} = -2.37$, $p = 0.02$; sevenfold: $t_{(24)} = -2.14$, $p = 0.04$). On the contrary, we found that the signal was negatively correlated with the fivefold and sevenfold models. This effect is a consequence of the fact that those models are themselves negatively correlated with the sixfold model (fivefold vs sixfold: $r = -0.49$, $p = 0.007$; sevenfold vs sixfold: $r = -0.51$, $p = 0.006$). Second, we verified that the sixfold modulation was not present before learning ($t_{(24)} = -0.55$, $p = 0.58$). The difference between prelearning and postlearning effects was marginally significant ($t_{(24)} = 1.89$, CI = [−0.01, 0.26], $p = 0.07$). Finally, motivated by theoretical and simulation works showing that grid cell activity can be used to estimate distance between spatial locations to subserve navigation and path integration (Bush et al., 2015), we applied model-based RSA, looking for distance effect (see above) in the entorhinal cortex. Although this region did not emerge from our previous distance-dependent, model-based, whole-brain searchlight RSA, we reasoned that, in light of the well known signal drop in the medial temporal lobe (Schmidt et al., 2005; Bellgowan et al., 2006; Olman et al., 2009), such a distance effect could have been overlooked by our strict whole-brain correction statistical threshold. Indeed, we did find a weak, but significant, distance effect in this area ($t_{(24)} = 2.14$, $p = 0.02$, one-tailed *t* test) that was not present before learning ($t_{(24)} = 0.48$, $p = 0.62$, one-tailed *t* test).

Reconstructing the semantic space

Finally, using multidimensional scaling, we wanted to test, also following Garvert et al. (2017), whether the neural activity in our two key regions (mPFC and entorhinal cortex) allowed faithful reconstruction of the relative positions of concepts in our novel semantic space (see Materials and Methods). Although before learning the reconstructed bidimensional space did not bear any similarity to the real arrangement of our concepts, this emerged after learning in both areas (Fig. 4*B*,*C*).

Discussion

In this experiment, we used fMRI to test whether spatially tuned representational codes, known to support spatial navigation in mammals, also subtend navigation of a semantic space in humans. Using a categorical learning paradigm of artificial stimuli that allowed us to master the precise metrics of a novel semantic space, and analyzing BOLD activity using fMRI adaptation and RSA, we found evidence of both distance and directional codes in regions typically involved in spatial navigation: the medial prefrontal cortex and the right entorhinal cortex. These results support the idea that the same neuronal machinery that is involved in navigating the physical environment in mammals is also recruited when humans navigate a categorical, labeled, conceptual space.

A previous study by Constantinescu et al. (2016) reported a direction-based modulation of BOLD signal in the mPFC and EHC while participants processed novel bird shapes that varied in their neck:legs ratio. This was taken as evidence of an internal "cognitive map" of conceptual knowledge, supporting navigation through concepts as they were locations in the physical space. Although the notion of "concept" in that seminal study was taken to indicate a conjunctive representation of two object characteristics (the length of the neck and legs), it remained silent on two key features that define semantic representations in humans: they refer to *categories* of objects or events (thus discrete entities), and they are accessible using *symbols* (words). In the current work, we aimed at overcoming these limitations. Similar to Constantinescu et al. (2016), we worked on a highly controlled 2D stimulus set, but contrary to them, we trained subjects to parse the 2D object space into four categories using words, engaging them in a symbolic categorization task. By associating more object exemplars to a single category name, participants were constructing true semantic representations, meeting all the definitional criteria of human-like concepts: that of originating from an arbitrary conjunction of multiple features [similar to Constantinescu et al. (2016), but here extended to the multisensory domain]; that of being categorical (thus defining more broad regions useful for generalization, as we demonstrated on the last training day); and that of being labeled with words.

Map-like code in the mPFC: a metric for distance

We first applied a series of univariate and multivariate analyses to find what brain regions represent the *distances* between the novel concepts, as the first ingredient of any map is to proportionally reflect the existing distances between locations in the space it

replicates. We found strong evidence of distance coding in the mPFC, where spatially tuned signals have been previously recorded in humans using both electrophysiology (Jacobs et al., 2013) and fMRI (Doeller et al., 2010; Constantinescu et al., 2016). Because the mPFC has been previously associated with a great variety of cognitive functions different from spatial cognition [see Stalnaker et al. (2015) for an exhaustive review], it has been proposed that this region actually represents the "task space," an internal model of the possible states in which an individual could be while performing a task (Stalnaker et al., 2015; Schuck et al., 2016; Behrens et al., 2018). This proposal not only is consistent with findings in the spatial navigation literature (the physical space would be the behaviorally relevant task space during navigation) but would also predict the recruitment of the same representational codes for nonspatial information, such as the concepts in our experiment: our task space, indeed, was a categorical and labeled multisensory concept space, where different concepts could vary along one or two sensory features that determined their proximities or distances.

Map-like code in the EHC: a metric for direction

We also investigated another important ingredient of any navigation device: the representation of the directions between visited locations. We asked whether we could find evidence for a directional code underlying navigation between concepts. Whereas previous studies relied on quadrature filter procedures (Doeller et al., 2010; Constantinescu et al., 2016; Nau et al., 2018) to reveal hexadirectional modulation of the BOLD signal, here we combined the multivariate approach first adopted by Bellmund et al. (2016) together with model-based RSA (Kriegeskorte et al., 2008), constructing a potentially more flexible method. Stensola et al. (2012) showed step-like discontinuity in grid orientation in rats' entorhinal cortex, sufficient to define discrete grid modules that are differently oriented. By assuming that the same organization would be preserved in humans, we reasoned, following Bellmund et al. (2016) and Bao et al. (2019), that multivariate analyses, capitalizing on the distributed activity across voxels, would be sensitive to the different preferred orientations of the grid modules also when moving in a conceptual space whose dimensions are, in psychological terms, large enough (as demonstrated by our psychophysical validation) to define navigable regions (the concepts or categories). Although we could not find evidence for a directional code in the mPFC, through whole-brain searchlight we found it in the right EHC, the brain region where grid cells in rats had been originally described (Hafting et al., 2005) and where human fMRI showed a directional representation during spatial (Doeller et al., 2010) and imagined (Bellmund et al., 2016) navigation. Although caution should be adopted in making inferences on neural codes following fMRI analyses, our study suggests that when subjects are engaged in a task emphasizing comparative judgments between subsequent stimuli that can be conceived as movements between regions of a bidimensional abstract space, the EHC displays a directional code similar to the one that supports navigation in the bidimensional physical space. Interestingly, in the same area, we also found a weak, but significant, distance code. Such a distance-dependent representation would be indeed predicted by the idea (supported by computational models) that directionally tuned grid cells support path integration and, thus, the representation of distances between locations (Howard et al., 2014; Bush et al., 2015). It is interesting to note that we did not observe the specular pattern in the mPFC: although we found a very strong modulation of its signal as a function of distance, we did not find evidence of directional coding. A plausible explanation is that, in our design, the two regions—mPFC and EHC—both support the representation of a map of concepts useful to solve the task, but by playing different roles: in light of the well known connectivity patterns between the hippocampal formation and the mPFC (Preston and Eichenbaum, 2013), the former might inform the latter by using a grid code. In this picture, the distance code recovered in the mPFC might reflect the output of the computations happening at a lower level in the EHC. Future studies should try to address this issue with more specific experimental designs and measures with finer temporal resolution.

Possible limitations of the present study

There is actually a second and potentially more parsimonious reason why we did not observe a directional modulation of the mPFC signal, which could also explain the weak directional result in the EHC and its absence in the mPFC [contrary, for instance, to Constantinescu et al. (2016)]: a subsampling of the possible movement directions within our 2D conceptual space. Because we could only sample eight directions, we might lack enough information to properly capture the grid signal compared with other experiments (Constantinescu et al., 2016) where, thanks to the availability of a continuous space, many different directions could be tested (at the cost of making less plausible a generalization to *discrete* human-like conceptual spaces). Moreover, due to the same subsampling issue, our sixfold symmetry model would also be compatible with a twofold rotational symmetry. Whereas a sixfold symmetry may be seen as readily deriving from the known hexagonal arrangement of the tuning functions of entorhinal grid cells (Hafting et al., 2005; Doeller et al., 2010; Jacobs et al., 2013; Bellmund et al., 2016; Constantinescu et al., 2016; Julian et al., 2018; Nau et al., 2018), a twofold rotational symmetry would correspond to a population of neurons tuned only to a given direction, φ, and to its opposite, $\varphi + 180°$, which to date has been reported only once, with important differences compared with the current case: they were not found in humans but in rodents, and not in the entorhinal but in the retrosplenial cortex (Jacob et al., 2017). However, even in the case in which our results actually derive from the activity of neuronal populations coding the stimulus space with a twofold symmetry, they would still represent evidence for a directional coding.

Additionally, we observed a significant *negative* correlation with two control periodicities (fivefold and sevenfold) that potentially originated by the anticorrelation existing a priori between these control models and the sixfold model. An alternative, complex, and (in our opinion) implausible explanation would indicate this result as specifically driven by a neural code tuned to the opposite (hence the negative correlation) of both a fivefold and a sevenfold symmetry at the same time, leading to an interpretation of the *positive* correlation with the sixfold model as a spurious consequence of the a priori anticorrelation between the tested models. We believe this interpretation would be biologically hard to sustain. In light of the well known presence of grid cells in the entorhinal cortex, tuned to a sixfold periodicity, we believe that our results are in line with the more parsimonious interpretation of a hexadirectional coding.

Conclusions

To conclude, in humans, a symbol-dependent categorical format of representations defines behaviorally relevant regions of the knowledge space that we typically refer to as "concepts." As human cognition critically depends on language, encoding the relationships between its units (the meaning of the words) is

fundamental for generalization, abstractions, and inferences, the key elements of human flexible cognition. Our results, by showing a modulation of the mPFC and the EHC (previously associated with spatial navigation) as a function of traveled distance and direction in the *semantic* space, may be seen as a novel example of "cortical recycling" (Dehaene and Cohen, 2007): brain regions holding specific coding schemes that evolved, in lower-level animals, to represent spatial relationships between spatial locations in humans are reused—or "recycled"—to encode relationships between words and concepts in an internal abstract representational space, akin to the previously hypothesized "cognitive map" (Tolman, 1948; O'Keefe and Nadel, 1978).

References

Bao X, Gjorgieva E, Shanahan LK, Howard JD, Kahnt T, Gottfried JA (2019) Grid-like neural representations support olfactory navigation of a two-dimensional odor space. Neuron 102:1066–1075.e5.

Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018) What is a cognitive map? Organizing knowledge for flexible behavior. Neuron 100:490–509.

Bellgowan PSF, Bandettini PA, Van Gelderen P, Martin A, Bodurka J (2006) Improved BOLD detection in the medial temporal region using parallel imaging and voxel volume reduction. NeuroImage 25:1244–1251.

Bellmund JLS, Deuker L, Navarro Schröder T, Doeller CF (2016) Grid-cell representations in mental simulation. eLife 5:e17089.

Bellmund JLS, Gardenfors P, Doeller CF (2018) Navigating cognition: spatial codes for human thinking. Science 362:6415.

Borghesani V, Piazza M (2017) The neuro-cognitive representations of symbols: the case of concrete words. Neuropsych 105:4–17.

Bush D, Barry C, Manson D, Burgess N (2015) Using grid cells for navigation. Neuron 87:507–520.

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu Y, Abdi H, Haxby JV (2012) The representation of biological classes in the human brain. J Neurosci 32:2608–2618.

Constantinescu AO, O'Reilly JX, Behrens TEJ (2016) Organizing conceptual knowledge in humans with a gridlike code. Science 352:1464–1468.

Dehaene S, Cohen L (2007) Cultural recycling of cortical maps. Neuron 56:384–398.

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

Ekstrom AD, Kahana MJ, Caplan JB, Fields TA, Isham EA, Newman EL, Fried I (2003) Cellular networks underlying human spatial navigation. Nature 425:184–188.

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558.

Gärdenfors P (2000) Conceptual spaces: The geometry of thought. Cambridge, MA: MIT.

Garvert, MM, Dolan RJ, Behrens TEJ (2017) A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife pii: e17086.

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436:801–806.

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

Henson RN, Shallice T, Gorno-Tempini ML, Dolan RJ (2002) Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex 12:178–186.

Henson RN (2016) Repetition suppression to faces in the fusiform face area: A personal and dynamic journey. Cortex 80:174–184.

Howard LR, Javadi AH, Yu Y, Mill RD, Morrison LC, Knight R, Loftus MM, Staskute L, Spiers HJ (2014) The hippocampus and entorhinal cortex encode the path and Euclidean distances to goals during navigation. Curr Biol 24:1331–1340.

Jacobs J, Weidemann CT, Miller JF, Solway A, Burke JF, Wei XX, Suthana N, Sperling MR, Sharan AD, Fried I, Kahana MJ (2013) Direct recordings of grid-like neuronal activity in human spatial navigation. Nat Neurosci 16:1188–1190.

Jacob PY, Casali G, Spieser L, Page H, Overington D, Jeffery K (2017) An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. Nat Neurosci 20:173–175.

Julian JB, Keinath AT, Frazzetta G, Epstein RA (2018) Human entorhinal cortex represents visual space using a boundary-anchored grid. Nat Neurosci 21:191–194.

Kaplan R, Friston KJ (2019) Entorhinal transformations in abstract frames of reference. PLoS Biol 17:e3000230.

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

Nau M, Navarro Schröder T, Bellmund JLS, Doeller CF (2018) Hexadirectional coding of visual space in human entorhinal cortex. Nat Neurosci 21:188–190.

O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res 34:171–175.

O'Keefe J, Nadel L (1978) The hippocampus as a cognitive map. Oxford: Clarendon.

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

Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in MATLAB/GNU octave. Front Neuroinform 10:27.

Piazza M, Izard V, Pinel P, Le Bihan D, Dehaene S (2004) Neuron 44:547–555.

Preston AR, Eichenbaum H (2013) Interplay of hippocampus and prefrontal cortex in memory. Curr Biol 23:R764–R773.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435: 1102–1107.

Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human orbitofrontal cortex represents a cognitive map of state space. Neuron 91:1402–1412.

Schmidt CF, Degonda N, Luechinger R, Henke K, Boesiger P (2005) Sensitivity-encoded (SENSE) echo planar fMRI at 3T in the medial temporal lobe. NeuroImage 25:625–641.

Stalnaker TA, Cooch NK, Schoenbaum G (2015) What the orbitofrontal cortex does not do. Nat Neurosci 18:620–627.

Stensola H, Stensola T, Solstad T, Froland K, Moser MB, Moser EI (2012) The entorhinal grid map is discretized. Nature 492:72–78.

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55: 189–208.

Watson AB, Pelli DG (1983) QUEST: a Bayesian adaptive psychometric method. Percept Psychophys 33:113–120.