The Hippocampus Maps Concept Space, Not Feature Space

Stephanie Theves, Guillén Fernández, and Christian F. Doeller

1Max-Planck-Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany, 2Donders Institute for Brain, Cognition, and Behaviour, Radboud University and Radboud University Medical Center, 6525 EN Nijmegen, The Netherlands, and 3Kavli Institute for Systems Neuroscience, Centre for Neural Computation, The Egil and Pauline Braathen and Fred Kavli Centre for Cortical Microcircuits, NTNU, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

The hippocampal formation encodes maps of space and a key question in neuroscience is whether its spatial coding principles also provide a universal metric for the organization of nonspatial, conceptual information. Previous work demonstrated directional coding during navigation through a continuous stimulus feature space as well as mapping of distances in a feature space that was relevant for concept learning. Here we provide the first unambiguous evidence for a hippocampal representation of the actual concept space, by showing that the hippocampal distance signal selectively reflects the mapping of conceptually relevant rather than of all feature dimensions. During fMRI scanning of 32 human participants (21 females), we presented everyday objects, which had beforehand been associated with specific values on three continuous feature dimensions. Crucially, only two dimensions were relevant to prior concept learning. We find that hippocampal responses to the objects reflect their relative distances in a space defined along conceptually relevant dimensions compared with distances in a space defined along all feature dimensions. These findings suggest that the hippocampus supports knowledge acquisition by dynamically encoding information in a space spanned along dimensions that are relevant in relation to define concepts.

Key words: conceptual knowledge; fMRI; hippocampus; learning; spatial coding

Significance Statement

How are neural representations of conceptual knowledge organized, such that humans are able to infer never experienced relations or categorize new exemplars? Map-like representations as supported by the hippocampal formation to encode physical space during navigation have been suggested as a suitable format. Here we provide the first evidence for a hippocampal representation of a conceptual space compared with a general feature-based space.

Introduction

The role of the hippocampus in concept learning is subject to debate (Knowlton and Squire, 1993; Zaki, 2004; Kumaran, 2012). Concepts are organizing structures that define how contents are related to each other and can be used to transfer meaning to novel input (Smith and Medin, 1981; Kemp, 2012). Their formation thus inherently depends on generalization over, and integration of experiences. Thus, a role of the hippocampus in generalization seemed considerable because of its roles in binding elements into spatial and episodic context (Davachi et al., 2003; Davachi, 2006; Ranganath, 2010; Komorowski et al., 2013) as well as integration of information over episodes (Davis et al., 2012; Collin et al., 2015; Milivojevic and Doeller, 2013; Milivojevic et al., 2015; Schlichting et al., 2015; beyond the episodic and spatial domain: Mack et al., 2016; Theves et al., 2019). Specifically, previous studies reported an involvement of the hippocampus in categorization (Nomura et al., 2007; Zeithamova et al., 2008; Davis et al., 2012; Mack et al., 2013; Seger et al., 2015; Kim et al., 2018). How specific this involvement is with regard to the conceptual aspect of the task and how the hippocampus, as opposed to other brain regions, supports the acquisition of conceptual knowledge remained unclear. A recent proposal is that map-like organization of new information by the hippocampal-entorhinal system similar to mental representations of space (O’Keefe and Dostrovsky, 1971; Hafting et al., 2005; Morgan et al., 2011; Howard et al., 2014; Horner et al., 2016), might be specifically suited to explain inference of not directly experienced relations (c.f. inferring shortcuts during navigation) (Behrens et al., 2018), as well...
as the transfer of meaning to novel information via localization of new input in the conceptual map. However, spatial coding principles have yet never been unambiguously linked to concept learning. So far, evidence for spatial coding in nonspatial domains has been limited to nonspatial feature dimensions without a direct link to conceptual relevance: Electrophysiological recordings in rodents demonstrated the involvement of place and grid cells in coding sound frequency during an auditory discrimination task (Aronov et al., 2017), and human fMRI studies showed both a directional, grid cell-like signal in entorhinal cortex during active navigation through a stimulus feature space (Constantinescu et al., 2016), as well as a hippocampal representation of distances in multidimensional feature space that was relevant to concept learning (Theves et al., 2019). Specifically, in the latter study, participants acquired a concept of two stimulus categories, which was defined in two-dimensional space along the feature dimensions of the stimuli, via a categorization task. Hippocampal representations of distances in concept space were measured via responses to passively viewed objects that had before been associated with specific stimuli (i.e., positions in concept space). As the associated stimuli only included features that defined the space of the concept, the important question remained whether the hippocampus maps the objects according to all feature dimensions of their associated stimuli (feature space mapping) or specifically for the purpose of concept learning (concept space mapping). Here we aim to distinguish between these two accounts by orthogonally manipulating conceptual and feature-based relationships between objects during learning. In sum, we show that the hippocampal responses to objects reflect the two-dimensional distances between objects that emerge from their positions in concept space, compared with distances that emerge in a space including an integration of the conceptually irrelevant feature dimension.

Materials and Methods

Experimental design and subject details
Thirty-two healthy students (mean age: 23 ± 3 years; 21 females) from the Radboud University campus participated in this study. All participants were right handed and had normal or corrected-to-normal vision. All participants gave written informed consent and were financially compensated for participation. The study was approved by the local ethics committee (CMO Arnhem-Nijmegen, The Netherlands).

Design
In a learning phase, participants acquired a novel concept of two abstract stimulus categories which was defined within a two-dimensional space along two of three stimulus feature dimensions with the diagonal through the two-dimensional space serving as category boundary (Fig. 1). The third stimulus feature was irrelevant to categorization and did not contribute to the concept space. Participants further learned to associate six everyday objects with six specific abstract stimuli. Here, associated abstract stimuli had to be memorized precisely in all three feature dimensions. Now critically, before and immediately after the learning phase, the everyday objects were presented in the MRI scanner to test whether hippocampal responses to the objects correspond specifically to their two-dimensional conceptual distances rather than to their three-dimensional distances emerging in the full feature space. To distinguish between concept- and feature-based mapping, objects were position such that their two- and three-dimensional distance relations were uncorrelated (Pearson’s r = 0.1).

Stimuli
The experiment involved the following two sets of stimuli: everyday objects (generated with the video game Sims; www.thesims3.com) and abstract stimuli (generated via MATLAB 2014a; Fig. 1B, associations). The abstract stimuli varied along the following three stimulus feature dimensions: opacity, frequency of dots, and frequency of stripes. On each dimension, abstract stimuli could vary along 10 steps, resulting in a total stimulus space of 1000 feature combinations. Step sizes on each of the three feature dimensions were evaluated psychophysically before the experiment to assure comparable discriminability of all three dimensions.

Procedures
The study took place on 1 d. A learning phase in which the conceptual context of six objects was acquired over the course of four tasks (details below), was preceded and followed by object-viewing blocks (OVB) in the scanner (Fig. 1). During the OVB, participants performed a target-object detection task (orthogonal to any conceptual content) to assure attention to the stimuli.

Object viewing block
Images of seven objects, of which six where used in the learning phase and one served as a catch-object, were presented in a pseudorandomized sequence with a stimulus duration of 1 s and interstimulus intervals of 3.5, 5, and 6.5 s (33.3% each). Participants were instructed to indicate for each object whether it is a trampoline (i.e., a catch object) or not, using a button box (buttons counterbalanced across participants). The task included 246 trials (236 for participant 10) with a catch trial rate of 12%. Each object was presented equally often.

The learning phase comprised four tasks in the following order: associative learning, 3D reconstruction, categorization, and navigating concept space, all before the post-learning OVB. The post-learning OVB was followed by a final 3D recall test (Fig. 1).

Associative learning
Associations between the six objects presented in the OVB and specific abstract stimuli had to be learned in alternating encoding and test blocks. The assignment of objects to abstract stimuli was randomized across participants, such that measuring hippocampal responses to the objects during scanning enabled us to read out their conceptual distances rather than visual or semantic similarities. In the encoding blocks, objects were presented next to their corresponding abstract stimulus and participants were instructed to memorize the presented pairs. Participants were told that they will need to memorize the associations in all their features throughout the entire experiment. The presentation order of the six pairs was pseudorandomized with each object/stimulus being equally often presented on the left/right position of the screen. Pairs were presented for 2 s on the screen and each pair was shown three times per encoding block. Each encoding block was followed by a test block in which the object is presented in the center of the screen along with the six abstract stimuli displayed (in a randomized order) below the object. Every association was tested once in blocks 1–6 and twice from block 7 onward in a randomized order. Participants selected the abstract stimulus associated with the presented object via key press (1–6) and received feedback (500 ms) on whether the choice was correct. Participants underwent at least eight encoding and test blocks (i.e., 60 test trials), and beyond that were trained until exceeding 90% accuracy over all previous test trials. An upper limit of 168 test trials was set because of the limited time between the prescheduled fMRI sessions.

3D reconstruction
Encoding and recognition of associations was followed by a free recall. Participants were instructed to precisely recall the abstract stimulus associated with a presented object, and subsequently adjust a start stimulus in all three feature dimensions until it matches the associated stimulus. A trial could only be completed by adjusting all three dimensions correctly. Each of the six abstract stimuli had to be reconstructed once. Dimensions were upregulated and downregulated using six adjacent keys.
Categorization
Participants were instructed to categorize abstract stimuli (see Stimuli) into two categories (A and B symbols), based on the relation of the opacity and dot frequency of a stimulus, whereas stripe frequency was unpredictable of category membership. Categories were, unbeknown to the participant, delineated via the diagonal (Fig. 1, dashed line) through a two-dimensional space spanned by the two relevant feature dimensions (2D concept space). As training stimuli, we selected for each participant a subset of 720 stimuli from the total stimulus space (1000 possible 3D feature combinations), including all possible off-diagonal combinations of opacity and dot frequency with randomly selected stripe frequency values. Stimuli were presented in a randomized sequence. In each trial, one abstract stimulus was presented in the center of the screen, and its category had to be selected via key press. Participants were given a maximum of 6 s to respond, and each response was followed by feedback (500 ms). Categorization training included at least 300 trials and afterward stopped when accuracy exceeded 85% across all previous trials or the maximal number of 720 trials (set because of time constraints between fMRI sessions) was reached. Instructions did not include any indications of a spatial rule.

We did not define an absolute performance criterion for “associative learning” and “categorization” since we expected high across-subject variance in both tasks (based on previous and pilot work), and time was constrained given the learning phase taking place in between two scheduled fMRI sessions. We also did not opt for a fixed number of trials to avoid unnecessary training in one task when a participant would have needed more trials to achieve high performance in the other task. Instead, with the present combination of trial limit and performance criterion, we intended to optimize the division of training time between these two tasks on an individual-subject level. Importantly, the term “criterion” does not refer to the exclusion criterion (this instead was based on the actual chance level (16.6% in associative learning, 50% in categorization).
Navigating 2D concept space

In each trial, an abstract stimulus (selected from the total pool of 1000 possible 3D combinations) was presented and participants were instructed to "collect" an object of a certain category (i.e., "Collect an A-object") by editing the two feature dimensions that were relevant for category membership (using four adjacent keys). Thus, they had to combine their knowledge of the specific object–stimulus pairs with categori- cal rule knowledge. A trial was completed when they navigated to one of the object "locations" that met the category-specific ratio of opacity and dot frequency and the collected object appeared together with its associated 3D stimulus on the screen. The third, conceptually irrelevant dimension had not to be adjusted by the participants. Instead, it was randomized across trials and remained constant within a trial. Once a participant correctly adjusted opacity and dot frequency to one of the required object locations, also the third feature switched to the respective value, serving as a reminder of the 3D associations relevant in subsequent tasks. There were 30 different start positions: 15 in category A, 15 in category B. Targets (Collect A-object or Collect B-object) were assigned pseudorandomly to the start positions, such that in half of the trials participants had to change the category field, while in the other half they did not. The task comprised 60 trials. Participants were instructed to collect each of the six objects at least seven times. The rationale behind the task was to familiarize participants with the conceptual context of the objects. Importantly, no distance relationships between the objects were introduced through this process, because participants did not navigate between the locations of the objects but started from random positions in the feature space.

3D recall test (subsequent to postlearning OVB)

We wanted to assure, that adjusting only the conceptually relevant dimensions during navigation did not result in a better memory of these two over the third feature of the six object-associated stimuli. Thus, on being cued by an object, participants had to adjust the three features of a start stimulus to match the stimulus associated with the object and eventually confirm their choice to enter the next trial. Each abstract stimulus had to be constructed four times. This allowed us to compare error rates in recall accuracy of the three feature dimensions. One participant did not conduct this task.

All tasks were conducted using Presentation 16.4 (NBS), except the 3D reconstruction, Navigation, and 3D recall tasks, which were programmed using Anaconda 2.7. (Python).

MRI methods

All images were acquired using a 3T PrismaFit MR scanner equipped with a 32-channel head coil (Siemens). A 4D multiband sequence (84 slices; multislice mode; interleaved; voxel size, 2 mm isotropic; TR = 1500 ms; TE = 28 ms; flip angle = 65°, acceleration factor PE= 2; FOV = 210 mm) was used for functional image acquisition. In addition, a structural T1 sequence (MPRAGE, 1 mm isotropic; TE = 3.03 ms; TR = 2300 ms; flip angle = 8°; FOV = 256 × 256 × 192 mm) was acquired. Separate magnitude and phase images were acquired to create a gradient field map (multiband sequence with voxel size of 3.5 × 3.5 × 2.0 mm; TR = 1020 ms; TE = 10 ms; flip angle = 45°).

Preprocessing of functional images was performed with FSL 5.0.9. Motion correction and high-pass filtering at 100 s was applied to the functional datasets. The following exclusion criteria for excessive motion were applied: mean absolute displacement >2 mm; or peak in absolute displacement >4 mm; mean ± SD of absolute displacement of the analyzed sample: 0.427 ± 0.205 mm (before) and 0.438 ± 0.199 mm (after). The FSL brain extraction toolbox was used to create a skull-stripped structural image. The structural scans were downsampled to 2 mm (matching the functional image resolution) and segmented into gray matter, white matter (WM), and CSF. Spatial smoothing (Gaussian) was performed at 3 mm. Mean intensity values at each time point were extracted for WM and used as nuisance regressors in the general linear model (GLM) analyses (see below). Structural images were registered to the MNI template. For each functional dataset (pre-learning, post-learning), the preprocessed mean image was registered to the individual structural scan and the MNI template. The coregistration parameters of the mean functional image were applied to all functional volumes.

Statistical analyses

fMRI data analysis: first level GLMs

All GLMs (GLM 1–2) included regressors accounting for catch trials and button presses as well as six motion parameters as covariates.

2D versus 3D: Distances between objects in the two-dimensional concept and three-dimensional feature space were modeled in the same GLM (GLM 1), using a stimulus onset regressor indicating the onset and duration of an object on the screen and two regressors being parametricaly weighted by the two-dimensional and three-dimensional distances between an object to the preceding object, respectively. Distances between objects in either space were calculated given the feature-based coordinates of the associated stimuli on the respectively relevant dimensions. Smaller distances were expected to result in lower signals, reflecting fMRI adaptation. We calculated the contrast between the two- and three-dimensional distance regressors (2D vs 3D contrast).

2D versus 2D(irrelevant) dimensionality control: If a potential difference in the 2D versus 3D contrast (GLM 1) would be merely because of a difference in dimensionality (i.e., a coding preference of the hippocamp for two dimensions) rather than because of a difference in conceptual relevance, two-dimensional distances in concept space (2Dxy) should not explain the hippocampal signal better than two-dimensional distances derived from a combination with the conceptually irrelevant z-axis (2Dxz, 2Dyz). Thus, we ran a GLM (GLM 2) with all three two-dimensional distance predictions as regressors (2Dxy, 2Dxz, 2Dyz) and contrasted the 2Dxy regressor against both alternative 2D regressors (2Dxz, 2Dyz). Resulting β-maps were transformed to MNI space to extract the average β value of each ROI for subsequent analysis.

fMRI data analysis: group-level analyses

First-level contrasts of the β estimates of the distance regressors were each averaged across all voxels within an ROI for each participant, and the distribution of these values was tested for significance (at α = 5%) using one-sample permutation t tests (Groppe, 2010) in which 1000 random permutations were computed to estimate the distribution of the null hypothesis. Correction for multiple comparisons for the number of spatial models tested (main analysis (GLM1): models 2D, 3D, post hoc dimensionality control analysis (GLM 2): models 2Dxy, 2Dxz, 2Dyz) were performed using the t-max method (Blair and Karniski, 1993). Because of clear directed predictions on the relations between fMRI adaption and distance (e.g., decreasing distance was supposed to be reflected in a higher fMRI adaptation, following the study by Theves et al., 2019), one-sided tests were applied. To test for effects on the whole-brain level, individual contrasts of the 2D versus 3D comparison were subjected to the second-level analysis. Cluster extend-based thresholding (z = 3.1, p = 0.05) was performed to correct for multiple comparisons.

ROI definition

For the hippocampal ROI mask, we thresholded probability maps from the Harvard-Oxford structural cortical atlas of the hippocampus at 50% probability.

Results

Behavior

Object detection task (fMRI session)

The six objects that were associated with abstract stimuli during learning plus an additional catch object were presented multiple times in a randomized sequence that was identical between the prelearning block and the postlearning block. Participants indicated via button press whether or not a presented object was the catch object. The task was performed with high accuracy (percentage of correct responses: pre-learning (mean ± SD): 98.28 ± 0.196%; post-learning: 98.018 ± 1.991%), indicating that participants paid attention to the objects.
Learning tasks

Associative learning. Associations between objects and abstract stimuli were studied in alternating encoding and test blocks. Participants performed between 60 and, maximally, 168 test trials (mean ± SD: 137.313 ± 40.689 trials), and within that range training was terminated on reaching accurate performance in 90% of all previous trials (see rationale for criterion in Procedures). The average final accuracy level was 87.315 ± 7.993% across all participants. Fifteen participants who did not fully reach the criterion within the trial limit were just short of 90% accuracy in the final trial (82.537 ± 9.710%). Thus, all participants exceeded chance level (i.e., 16.6%) by far.

3D reconstruction. The six object-to-abstract stimulus associations were each recalled once in a 3D reconstruction task, in which an object cue, the associated stimulus had to be reconstructed by adjusting all three feature dimensions to the correct value. Each trial ended on correct completion. An ANOVA comparing deviation of “different coordinates visited” from “required steps” across dimensions (F(11.8, p < 0.0001; post hoc paired t tests: x vs y: p = 0.013, t(31) = 2.739; x vs z: p = 0.001, t(31) = 4.391; y vs z: p = 0.115, t(31) = 1.608) shows that editing dimension z was accomplished with fewer unnecessary edits compared with dimension x and equally well relative to dimension y, indicating that the later conceptually irrelevant dimension z was initially encoded. The ultimate knowledge of all three dimensions at the end of learning and critical time of scanning is, however, appropriately captured by the final 3D recall test.

Categorization. Participants learned to categorize abstract stimuli within at least 300, but maximally 720 feedback-based trials (542.906 ± 205.428 trials). Within this range, training stopped when 85% of all previous trials had been classified correctly. Across all participants, the average accuracy was 82.892 ± 4.639. Sixteen participants narrowly missed 85% accuracy in their final trial (79.158 ± 3.488%). Thus, all participants performed considerably above chance-level (i.e., 50%).

Navigating 2D concept space. Categorical knowledge as well as knowledge about the six object associations had to be combined in a subsequent “navigation in concept space.” Here each trial required collecting an object of a certain category by adjusting the two conceptually relevant feature dimensions until a category-specific object location was reached. All objects were on average collected at least seven times.

3D recall test (postsansing). Subsequent to the final scanning session, recall accuracy of all three dimensions of the abstract stimuli was tested in a 3D reconstruction task that required participants to confirm their adjustments as soon as they considered them correct. Recall errors [deviation of reconstructed value from actual coordinate; opacity (10): 0.440 ± 0.472; frequency dots (y): 0.427 ± 0.474; frequency stripes (z): 0.701 ± 0.725; n = 31] did not differ across dimensions (F(2,62) = 2.44, p = 0.0932; n = 31; post hoc pairwise tests between relevant and irrelevant dimensions: x vs z: t(30) = −1.710, p = 0.109; y vs z: t(30) = −1.792, p = 0.091).

fMRI

Hippocampal signal reflects distances in a 2D concept space, not in a 3D feature space

We hypothesized that the hippocampus supports the formation of conceptual knowledge by organizing novel information in a space defined along conceptually relevant dimensions. Following the study by Thèves et al. (2019), we expected distances between objects in an abstract space defined along stimulus feature dimensions to be reflected in fMRI adaptation in which the distance to the preceding object would scale with the strength of the hippocampal response (smaller distances relate to higher similarity of the neural response pattern and thus in higher adaptation). The representation of feature-based distances reported in the study by Thèves et al. (2019) was shown to be specific to the hippocampus (no effects in whole brain or in ROI analyses on control regions: lateral occipital cortex, postcentral gyrus, entorhinal cortex). Here, we aim to further examine this hippocampal distance effect by probing whether hippocampal responses to the objects are explained specifically by distances between objects in a concept space (defined only along the two conceptually relevant stimulus feature dimensions) or by distance predictions derived from a space defined along all three stimulus dimensions in feature space. We found that hippocampal adaptation significantly scaled with distances between objects in the two-dimensional concept space, but not with distances derived from the full three-dimensional feature space (2D: t(31) = 3.090, p = 0.003; 3D: t(31) = −1.434, p = 0.916; corrected for multiple comparisons; see Materials and Methods). The two-dimensional conceptual distances also explain the hippocampal response significantly better than the three-dimensional feature-based distances (contrast 2D vs 3D: t(31) = 3.163, p = 0.001; Fig. 2B). We did not observe significant 2D versus 3D effects in other brain regions (whole-brain cluster-extend-based thresholding, z = 3.1, p = 0.05).

The 2D concept versus 3D feature space contrast does not reflect differences in dimensionality

If the better fit of the hippocampal response by the two-dimensional (vs the three-dimensional) distances would merely reflect a difference in dimensionality (i.e., a coding preference of the hippocampus for two dimensions) rather than a difference in conceptual relevance between both spaces, two-dimensional distances in concept space (xy) should not explain the hippocampal signal better than two-dimensional distance predictions derived from a combination with the conceptually irrelevant feature dimension [i.e., the opacity–dot frequency (xz) plane or stripe frequency–dot frequency (yz) plane]. Thus, we constructed a GLM with 2Dxy (concept space), 2Dxz, and 2Dyz as regressors. Only the 2D distance in concept space, but none of the alternative 2D models integrating z, predicts the hippocampal signal [2D(xy): t(31) = 2.678, p = 0.010; 2D(xz): t(31) = 0.268, p = 0.497; 2D(yz): t(31) = 1.569, p = 0.847; corrected for multiple comparisons]. The 2Dxy (concept space) also reveals a significantly stronger adaptation than the 2D controls (2Dxy vs 2Dxz: t(31) = 2.439, p = 0.015; 2Dxy vs 2Dyz: t(31) = 2.786, p = 0.004; corrected for multiple comparisons; Fig. 2C).

Discussion

For the first time, we demonstrate a direct link between concept learning and hippocampal representations of abstract spaces defined by nonspatial dimensions. While the hippocampus had before been shown to encode distances in a multidimensional feature space as a result of concept learning (Thèves et al., 2019), we here intended to discriminate whether this representation reflects the complete feature space or specifically the space embedding the concept. In our design, object relationships could be described in a two-dimensional concept space defined along only conceptually relevant stimulus dimensions and in a three-dimensional feature space defined along all stimulus dimensions. We found that hippocampal representations of objects encountered during prior concept learning reflected their concept- and not their feature-based distances. This effect could not be attributed to the difference in dimensionality between concept and
feature space predictions and a potential coding preference of the hippocampus for two dimensions, as the two-dimensional distances in concept space also explained the hippocampal signal significantly better than alternative two-dimensional distances derived from combinations with the conceptually irrelevant feature dimension. Thus, the hippocampal signal reflects only a representation of distances in a space spanned by the dimensions that were relevant in relation to each other to define the concept, while the mnemonically, but not conceptually, relevant third dimension was not integrated in a multidimensional representation. In sum, we show that the hippocampus organizes new information in a map-like representation in support of concept learning.

First, this suggests that during concept learning, the hippocampus actively organizes new information in a multidimensional space according to conceptually relevant dimensions, and not according to any perceptually present information. The notion that hippocampal spatial codes might not involve incidental sensory information is also in accordance with recent investigations in rodents (Aronov et al., 2017). In the context of the current study, it should further be noted that the better fit of concept-based distances to the hippocampal signal also speaks against the distance representation being a secondary effect, reflecting the similarity of the sensory input to the hippocampus during potential pattern completion to the associated three-dimensional stimuli.

The present effect further distinguishes between conceptual and general task relevance: all three feature dimensions are task relevant with respect to the mnemonic component of the learning phase. Specifically, concept learning (i.e., categorization) requires setting two feature dimensions in relation to each other, making a map-like representation that integrates both dimensions advantageous. Thus, regarding the question of whether spatial codes in the hippocampus are domain general, it is conceivable that the hippocampus organizes information along arbitrary dimensions (spatial or abstract) into map-like representations as long as the dimensions are relevant in relation to each other (Eichenbaum, 2004; navigation in or representation of 2D spaces; Constantinescu et al., 2016; Bao et al., 2019; Theves et al., 2019). Regarding the role of the hippocampus in concept learning, it is of interest to pursue whether the hippocampus actively organizes new information in a multidimensional space according to conceptually relevant dimensions, and not according to any perceptually present information. The notion that hippocampal spatial codes might not involve incidental sensory information is also in accordance with recent investigations in rodents (Aronov et al., 2017). In the context of the current study, it should further be noted that the better fit of concept-based distances to the hippocampal signal also speaks against the distance representation being a secondary effect, reflecting the similarity of the sensory input to the hippocampus during potential pattern completion to the associated three-dimensional stimuli.
learning in particular, hippocampal spatial codes can be consid-
erized a candidate mechanism that is specifically suited to address
the typically relational nature of conceptual knowledge (i.e.,
knowing which features and relations between them distinguish
different categories).

Despite evidence for three-dimensional spatial coding in the
hippocampus (Yartsev and Ulanovsky, 2013; Kim et al., 2017;
Porter et al., 2018; Wohlgemuth et al., 2018), it is currently
unknown whether the hippocampus would map a 3D concept
space if three dimensions were conceptually relevant. Accordingly,
one might speculate on whether a potential hippocampal prefer-
ence for two dimensions favors the 2D over the 3D model.

Critically, if this would be the only reason for the better fit of the
two-dimensional distances to the hippocampal signal, while there
are no differences regarding the integration of feature dimensions
in a combined representation, we would expect the two-dimen-
sional distances derived from combinations with the conceptually
irrelevant dimension to be encoded in the same way as the two-
dimensional conceptual distances and thus to likewise fit the hip-
campal response. We ruled out this alternative by showing that
two-dimensional distances in concept space (2Dxyz) explain hippo-
campal responses significantly better than two-dimensional dis-
tance predictions that were derived by combinations with the
irrelevant dimension [opacity–stripe frequency plane (2Dxz) or
dot frequency–stripe frequency plane (2Dyz)].

Further, we ensured that the better fit of the two-dimensional
contextual distances cannot be attributed to weak memory of
the conceptually irrelevant dimension. It should be noted that
initial encoding of 3D associations took place before the con-
ceptual relevance of two dimensions was introduced via the
categorization task and should thus not be affected by this
manipulation. To test memory as a result of all learning tasks at
the time of scanning, participants were required to reconstruct
the stimuli in all three feature dimensions after the postlearning
fMRI session. Importantly, this 3D recall test revealed very high
recall performance in all three dimensions (i.e., the average error
of 2.68° on average). It should be noted that the present results are not in contrast to the vast
body of literature demonstrating cortically distributed repres-
tations of concepts embedded in long-term semantic knowledge
(Martin, 2007, 2016; Binder and Desai, 2011; Ralph et al., 2017),
but propose a spatial code for the formation of concepts in the
hippocampus. Thus, while semantic information might ulti-
mitely be stored in neocortex, the hippocampus seems to crit-
ically support its acquisition (Kumaran, 2012; Elward and
Vargha-Khadem, 2018). For instance, although patients with de-
velopmental amnesia because of hippocampal atrophy can show
semantic memory comparable to that in control participants in
everyday life (Vargha-Khadem et al., 1997; potentially compen-
sated by direct cortical incorporation of new information into
existing representations over time), the learning of completely
new material (assumed to be critically supported by fast hippo-
campal processing of trial-unique stimuli) was shown to be ame-
liorated (Elward and Vargha-Khadem, 2018). Accordingly, a
hippocampal organization of new information into a map-like
format might support the acquisition of concepts, when fast
extraction of critical relations or structures and commonalities
across events is required. The present results suggest a role of the
hippocampus in the formation of cognitive spaces spanned by
relationally relevant feature dimensions, which provide sufficient
flexibility for inferential processes or transfer, and from which
more abstracted information (i.e., dichotomic category member-
ship responses classically observed in PFC rather than hippo-
campus; Freedman et al., 2001, 2003; Wallis and Miller, 2003; Seger
and Miller, 2010; Meyers et al., 2008; Roy et al., 2014) can be
derived and coded by other brain regions. Accordingly, the pres-
ent hippocampal representation reflects feature-based distances
in a space spanned by conceptually relevant dimensions, without
being driven by coarse category membership (post hoc analyses
including category membership to the GLM reported above
revealed that the 2D feature-based distance regressor remains
significant (t(31) = 1.983; p = 0.0185), while category membership
cannot explain the hippocampal signal (t(31) = −0.132,
p = 0.552). Instead, congruent with the conceptual nature of the
present feature-based representation, post hoc analyses reveal
that hippocampal responses to objects scales with the 2D dis-
tance of the objects to the category boundary (t(31) = 1.704;
p = 0.044; inter-object distance was included as a regressor and
remained significant: t(31) = 2.928, p = 0.0015). As categories had
been delineated via the diagonal through the two-dimensional
feature space, information about the boundary emerges only
from an integrated representation of conceptually relevant fea-
ture dimensions and can thus be considered further support for
the spatial format of the representation.

In sum, by demonstrating that the hippocampus encodes dis-
tances between points in a concept space, as opposed to a full
feature space, the present study provides critical evidence that hippocampal coding principles provide a suitable format to represent conceptual knowledge.

**References**


