bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Value shapes the structure of schematic representations in the medial prefrontal cortex

Philipp C. Paulus[a,b,✉], **Ian Charest**[c], **and Roland G. Benoit**[a]

[a]Max Planck Research Group: Adaptive Memory, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[b]International Max Planck Research School NeuroCom, Leipzig, Germany
[c]School of Psychology, University of Birmingham, Birmingham, UK

Adaptive cognition is fostered by knowledge about the structure and value of our environment. Here, we hypothesize that these two kinds of information are inherently intertwined as value-weighted schemas in the medial prefrontal cortex (mPFC). Schemas (e.g., of a social network) emerge by extracting commonalities across experiences and can be understood as graphs comprising nodes (e.g., people) and edges (e.g., their relationships). We sampled information about unique real-life environments (i.e., about personally familiar people and places) and probed the neural representations of their schemas with fMRI. Using representational similarity analysis, we show that the mPFC encodes both, the nodes and edges of the schemas. Critically, the strength of the edges is not only determined by experience and centrality of a given node but also by its value. We thus account for the involvement of the mPFC in seemingly disparate functions and suggest that valuation emerges naturally from encoded memory representations.

✉ *paulus@cbs.mpg.de*

Introduction

Our rich knowledge of the past allows us to readily make sense of the present. It also facilitates adaptive planning for the future, for example by supporting simulations of prospective events (1–5). Critically, these capacities are not exclusively dependent on individual memories of unique past experiences. Instead, they are also based on generalized knowledge about our environment that is derived from multiple experiences (e.g., about relationships between familiar people)(3, 6).

A type of such generalized knowledge structures are memory schemas (7–9). These representations of our environment can be understood as graphs comprising information about nodes (e.g., individual people) and their edges (i.e., their relationships) (8, 10–12). Schemas are formed by extracting commonalities across related events (7, 8, 13). They thereby reduce the complexity of our experience into simplified models of the world (e.g., about the people we know or about the locations we frequently visit) (8, 14). Such models, in turn, foster planning and facilitate adaptive decisions (15–17).

However, beyond a representation of the environment's structure, adaptive cognition also requires a representation of what's valuable within that environment (18). Here, we test the hypothesis that these two kinds of information are inherently intertwined in the rostral and ventral medial prefrontal cortex (mPFC) (19–22). As detailed below, this proposal accounts for the involvement of this region in two seemingly disparate functions: representing *memory schemas* and *affective value*.

Evidence from humans (23, 24) and rodents(20, 25) indicates a critical role for the mPFC in mediating *memory schemas* (26, 27). Activity patterns in this region have been shown to code for individual nodes of the environment, such as for familiar people (19, 28) and places(19, 29). However, it remains unclear whether the mPFC encodes representations of the nodes in isolation or whether these representations also entail information about their edges (i.e., their relationships to other nodes).

A largely independent line of research has associated the mPFC with the representation of *affective value* (21, 30–32). Activity in this region tracks the value of objects, people, or places that we currently perceive or imagine (19, 30, 33–35). Moreover, in humans, focal lesions disrupt value judgements (36, 37). The mPFC has thus been argued to represent value in a common currency that allows for flexible decision making in a wide range of contexts (30, 31, 38). Notably, evidence from human neuroimaging (1, 19, 33, 35, 39) and rodent single cell-recordings (20, 22) has shown that representations of memories and of affective value are supported by *overlapping* parts of the mPFC. We thus reconcile the common attribution of these functions by hypothesizing that the schemas encoded by this region are shaped by affective value.

Specifically, we propose that the mPFC encodes representations of individual nodes (e.g., individual familiar people) and that the representations also entail information about their edges (e.g., the overall associations between the people). Critically, we suggest that nodes that are more important for a person exhibit stronger edges.

We hypothesize that the importance of a given node is jointly determined by three features: Given that schemas build up with experience (8), we first expect that more familiar nodes should be more prominently embedded in the overall graph (33). Secondly, for the same reason, we expect stronger embedding of nodes that are more central to the respective environment (11). Finally, given the role of the mPFC in affect and valuation (21, 30), we propose the strength of the edges are also weighted by each node's affective value (19, 20, 40).

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The encoded schemas would thus emphasize connections of behaviorally relevant elements of the environment, reminiscent of the hippocampal weighting of rewarded locations (41, 42).

Here, we test this hypothesis by probing the neural representations of two distinct and individually unique schemas: about people's social networks and about places from their everyday environment. This allows us to examine whether the suggested coding principles generalize across these individual schemas. Participants provided names of people and places they personally know and arranged these names in circular arenas according to their associations (43) (Fig. 1 d). This allowed us to quantify the centrality of each exemplar (e.g., a person) to its respective schema (e.g., the social network). Participants also indicated their familiarity with each person and place (as an index of experience) and their liking of each of these exemplars (as an index of affective value). In a subsequent session, we measured their brain activity using functional magnetic resonance imaging (fMRI) while they imagined interacting with each person and being at each place. We took the ensuing activity patterns to assess the neural representations of the individual nodes and their edges using representational similarity analysis (RSA) (44).

First, we hypothesized that the mPFC encodes unique representations of the nodes that get reinstated during mental simulation (19, 28, 29). We thus predicted similar activity patterns to emerge in the mPFC whenever participants imagine the same person or place. Second, we hypothesized that the structure of neural activity patterns *between* nodes reflects the structure of their edges. We reasoned that pairs of nodes that are more strongly connected (i.e., that exhibit stronger edges) are encoded by more overlapping neuronal populations (1, 45–47). This, in turn, should be reflected in overall greater neural similarity for nodes with particularly strong edges. As a consequence, if more important nodes have stronger edges, they should also exhibit overall greater neural similarity.

In addition, we further gauge the regional specificity of such value-weighted schemas to the mPFC. Therefore, we also examine the posterior cingulate cortex and the hippocampus, two regions that have similarly been associated with memory (48–51) and valuation (30, 34, 52).

Results

The medial prefrontal cortex codes for the nodes of participants' real-life schemas. We first examined the hypothesis that the medial prefrontal cortex encodes representations of personally familiar people and places, i.e., the nodes of the respective schemas. Whenever we simulate an event involving a particular node, its representation should get reinstated in the mPFC. We thus took the ensuing fMRI activity patterns as proxies of their respective neural representations (19, 53) and examined their replicability using an RSA searchlight approach (radius = 8 mm, 4 voxels) (44).

In regions that encode the nodes of the schema, we predicted overall greater pattern similarity for simulations featuring the same node (same-item similarity) than for simulations featuring different nodes (different-item similarity) (54). Note that the different-item measure was only based on the similarity of activity patterns for nodes of the same kind (i.e., either people *or* places). This ensured that the results are not influenced by potential categorical differences in the representation of people versus places (Fig. 1 a, b) (19, 53).

Corroborating our previous finding (19), we obtained this effect in the mPFC. This region thus yielded replicable activity patterns that were specific to individual exemplars (Fig. 1 c, Supplement 1). Moreover, we also observed evidence for such replicable pattern reinstatement in a number of other brain regions that are typically engaged during the recollection of past memories and the simulation of prospective events (50, 55, 56). These regions included the posterior cingulate cortex (PCC), the precuneus and parts of the lateral parietal and temporal cortices. Notably, there was no evidence for pattern reinstatement in the hippocampus.

The data thus support our hypothesis that the mPFC encodes unique representations of individual nodes. In the following, we further examine the edges *between* nodes in the mPFC and PCC regions of interest (ROI) identified by this analysis. We also test for these effects in the hippocampus, even though this region showed no evidence of node coding.

Note that the subsequent analyses of the edges are based on different parts of the neural representational similarity matrix (RSM) than the ones used to determine node coding. Further, they are based on comparisons of model RSMs that are also independent of the node coding model. In the supplement, we provide complementary and consistent results based on anatomically defined masks (see Supplementary Tables 2 – 4; Supplementary Figure 1 and 2).

Centrality, experience, and affective value load on a common principal component that quantifies importance. We had hypothesized that a node's centrality, experience, and also its affective value contribute to its importance. These three features may thus share a common latent factor. First, to assess *centrality*, participants positioned the names of the people and places in circular arenas (Fig. 1 d). They were instructed to arrange nodes close together if they associate them strongly with each other and far apart if they do not (43). We calculated the centrality of each node as the sum of its inverse distances to all other nodes. Participants then arranged the people and places on continuous scales providing estimates of their familiarity with each node (as an index of *experience*) and of their liking (as an index of *affective value*). All three features were assessed separately for people and places.

To test whether centrality, experience, and affective value load on a common latent factor, we z-scored each vector

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

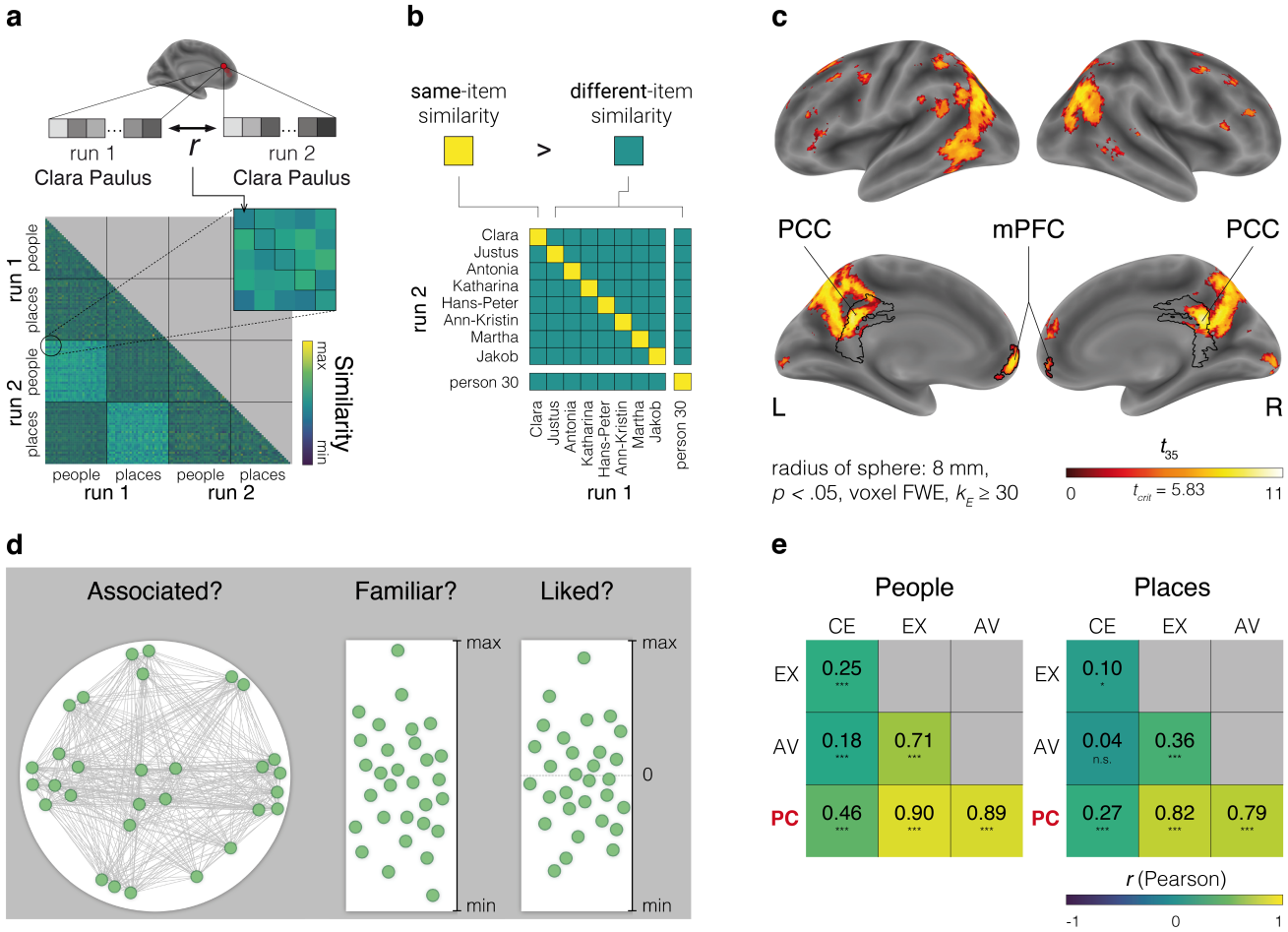

Fig. 1. Representations in the mPFC and PCC code for the nodes of participants' real-life schemas. a We examined whether the mPFC encodes representations of the nodes by testing for the replicability of activity patterns for the same people and places across the two functional runs. Each row and column of the representational similarity matrix corresponds to a single simulation trial. **b** Regions coding for the nodes should show more similar activity patterns for the repeated simulation of the same person or place (same-item similarity) than for simulations entailing different nodes of the same category (different-item similarity). **c** The searchlight analysis identified regions coding for the nodes of real-life schemas. These entailed the mPFC and PCC. **d** Participants arranged the familiar people and places on circular arenas according to their associations, thus providing a measure of centrality. Participants also provided measures of experience and affective value by indicating their familiarity with the nodes as well as their liking. **e** Centrality, experience, and affective value load on a common principal component as indicated by significant positive correlations (* - $p < .05$, *** - $p < .001$; one-tailed; $df = 35$). This component thus summarizes the importance of the nodes to the schema. mPFC = medial prefrontal cortex, PCC = posterior cingulate cortex, CE = centrality, EX = experience, AV = affective value, PC = principal component.

of values separately for each category (people, places) and within each participant. This approach prevents between-participant variance from influencing the factor solution. We then performed principal component analyses, separately for the people and places. The respective first principal component explained, across all participants, 61% of variance for people and 46% for places.

Critically, as predicted, both of these principal components exhibited significant positive correlations not only with experience and centrality but also with affective value (Fig. 1 e). We thus take them to quantify the importance of each individual node to its respective schema. In the next step, we used the individual importance values of the respective principal component to predict the structure of the schemas' edges.

The medial prefrontal cortex encodes the edges of value-weighted schemas. We had hypothesized that more important nodes – as indicated by the principal component – exhibit stronger edges. We had further reasoned that the strength of edges between nodes is reflected in their patterns of neural similarity, assuming that more strongly connected nodes are also encoded by more overlapping neuronal populations (1, 20, 57). We had thus predicted that more important nodes should exhibit overall greater neural similarity.

We tested this prediction by constructing models of the expected structure of representations in the mPFC. The models were based on the importance values derived from the respective principal component. Specifically, we predicted the similarity between any two nodes by the product of their respective principal component scores (i.e., importance values). Thus, we expected nodes that are more important to yield overall greater pattern similarity (Fig. 2 b).

We then determined the neural similarity structure in the mPFC, PCC, and in the hippocampus (Fig. 2 a). We con-

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

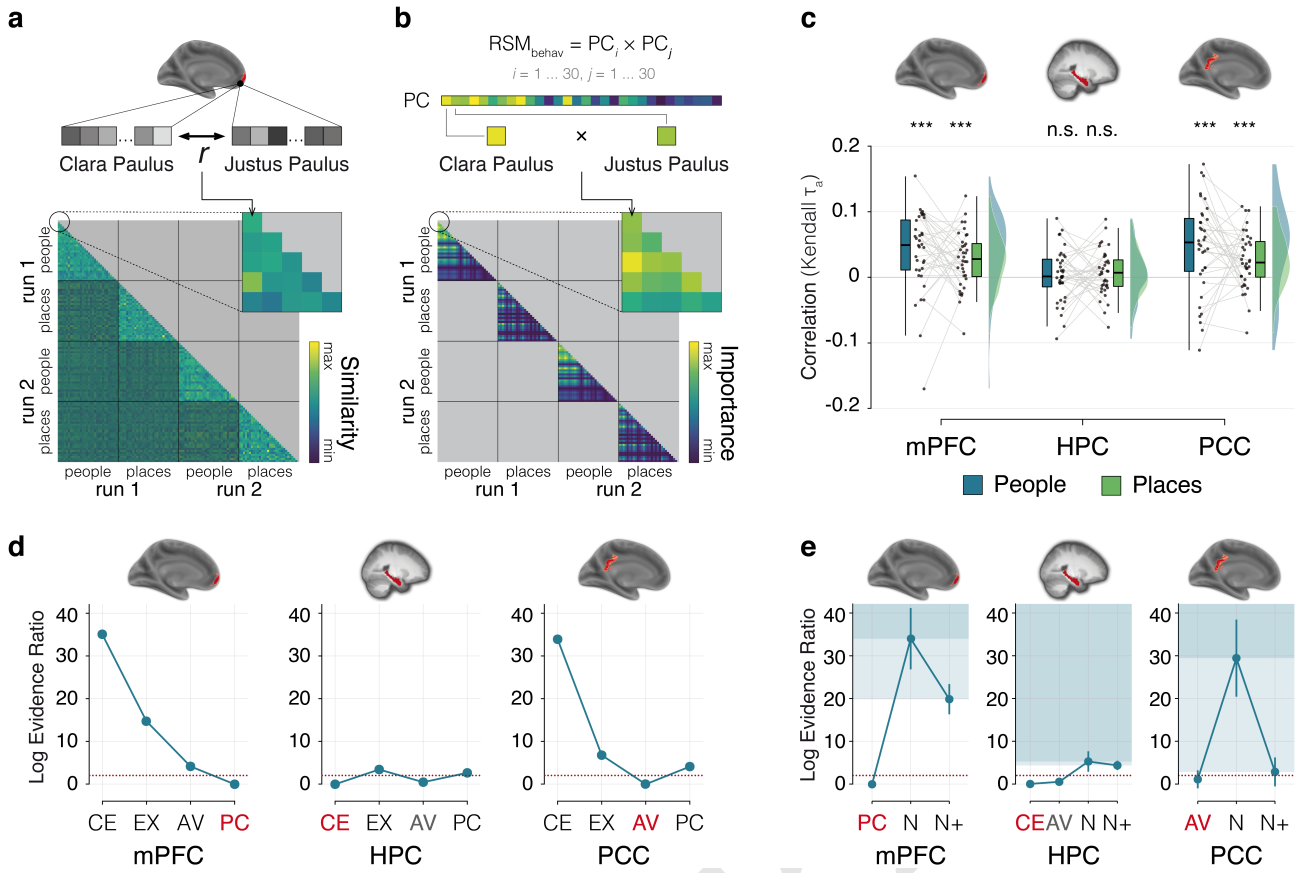

Fig. 2. Only the structure of representations in the mPFC is best accounted for by the principal component model that reflects importance. a Construction of the neural RSM. Each row and column of the matrix corresponds to a single simulation trial. In this analysis, we examine the similarity of activity patterns elicited by simulations of different people or places. **b** Construction of the model predictions. We predicted more similar representations for people and places with overall higher principal component scores, given that these more important exemplars should be more strongly embedded in their overall schema. To this end, we computed the combined importance of any two people or places from the product of their principal component scores. **c** Correlation of neural RSM and model prediction. Asterisks denote significant positive correlations as tested in a *t*-test on the *Fisher-z* transformed correlation coefficients (*** - $p < .001$; one-tailed; $df = 35$). **d** Comparisons of linear mixed models further support the hypothesis: only the structure of representations in the mPFC is best explained by the principal component. The figure displays Log Evidence Ratios (LER). Smaller values indicate better fit. By definition, the best model assumes a value of zero. The dotted red line demarks a relative LER difference of two, regarded as decisive. **e** Comparisons with noise null models. Data points depict the mean model performance in comparisons with 1,000 random noise models (N) and sorted noise models (N+), whiskers indicate the standard deviation of the model performance. Shaded areas correspond to the region > N (darker shade) and > N+ (lighter shade). mPFC = medial prefrontal cortex, HPC = hippocampus, PCC = posterior cingulate cortex, CE = centrality, EX = experience, AV = affective value, PC = principal component, N = random noise, N+ = sorted noise.

strained the broader cluster containing the PCC to the parts covering this region using an anatomical mask (58). We used an anatomical mask from the same atlas to examine the representational structure for the hippocampus. All analyses were conducted in subject space.

Finally, we tested for the correspondence between our prediction and the actual structure of neural representations by computing the correlation of the respective parts of the lower triangular vectors of both matrices (Fig. 2 a-c). This was done separately for people and places to examine whether the effect is present for either category. Using Kendall's τ_a as a conservative estimate (59), we indeed observed significant correlations in the mPFC for both people (*mean* τ_a = 0.039, tested with a Wilcoxon test, $W = 562$, $p < .001$, d = 0.63, one-tailed, due to a deviation from normality indicated by a Shapiro-Wilk test, $W = 0.92$, $p = 0.01$) and places (*mean* $\tau_a = 0.026$, $t(35) = 3.72$, $p < .001$, $d = 0.62$, one-tailed) - with no significant differences between the two ($t(35) = 1.04$, p = .307, $d = 0.17$, two-tailed).

Similarly, the correlations were also significant in the PCC for people (*mean* $\tau_a = 0.046$, $t(35) = 3.91$, $p < .001$, $d = 0.65$, one-tailed) and places (*mean* $\tau_a = 0.026$, $t(35) = 3.61$, $p <$ .001, $d = 0.60$, one-tailed), again with no significant differences between the two ($t(35) = 1.42$, $p = .165$, $d = 0.24$, two-tailed). However, the same analyses of the hippocampal data did not yield evidence for a match between the predicted and actual structure of representations (people: *mean* $\tau_a = 0.003$, $t(35) = 0.41$, $p = .341$, $d = 0.07$, one-tailed; places: (*mean* τ_a = 0.007, $t(35) = 1.15$, $p = .13$, $d = 0.19$, one-tailed).

We obtained qualitatively identical results in analyses based on purely anatomically defined ROIs (see Supplementary Table 2 and 3; Supplementary Fig. 1). The results are also in accordance with a whole-brain searchlight analysis (radius = 8 mm, 4 voxels) (see Supplementary Table 5). We thus show that representations in the mPFC generally align with the predicted structure of value-weighted schematic representations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, it remains to be determined whether importance is indeed the best model to account for the structure of representations in any of our ROIs.

The importance model accounts best for the structure of mPFC representations only. Does the structure of representations predicted from the principal component account best for the data or would any of the individual contributing features provide at least a comparable or even a better fit? If the mPFC does encode value-weighted schemas, we would expect the model based on the conglomerate index of importance to outperform models only based on centrality, experience, or affective value. Furthermore, we would expect some degree of regional specificity, i.e., that only representations in the mPFC are best accounted for by importance.

We formally tested these predictions by setting up alternative models that were solely based on either centrality, experience, or affective value. We then compared these models with the importance model that was based on the principal component. In brief, we set up linear mixed effects models to account for the structure of representations as a function of each of these individual features. In each of these models, we included a factor of category (people, places) and the maximum possible random effects structure that would converge across all models and regions of interest. We thus accounted for between participant variance by including a random intercept per participant and run, as well as random slopes for our fixed effects of category and the respective predictor (e.g., the principal component scores). We then performed model comparisons within each ROI to determine the model that best fits the neural similarity structure. The comparisons were based on Log Evidence Ratios (*LER*) derived from Akaike's Information Criterion (60). We regarded *LER* differences greater than two as decisive evidence for the better model (61).

Consistent with our hypothesis, in the mPFC, the principal component model accounted best for the data. It performed decisively better than affective value (*LER* = 4.19), experience (*LER* = 14.8) and centrality (*LER* = 35.16) (Fig. 2 d). The model parameters of this winning model entailed a significant main effect of category, reflecting overall higher neural pattern similarity for people than places ($\beta_{Category_place}$ = -0.026, *SE* = 0.008, χ^2 = 11.61, p < .001). Critically, they also included a significant positive parameter estimate for the principal component, indicating overall greater neural pattern similarity for nodes of greater importance ($\beta_{PrincipalComponent}$ = 0.048, *SE* = 0.012, χ^2 = 17.12, p < .001). Moreover, the main effect of the principal component did not interact with category ($\beta_{Category_place:PrincipalComponent}$ = -0.005, *SE* = 0.008, χ^2 = 0.36, p = .546).

By contrast, in both control regions, other models were better suited to account for the structure of representations. For the hippocampus, the model comparison yielded the best fit for centrality, though there was only a minimal advantage for this model over affective value (*LER* = 0.49). Notably, both models performed decisively better than the ones based on either the principal component (*LER* = 2.68) or experience (*LER* = 3.46). However, of the model parameters, only the main effect of category was significant, indicating overall higher pattern similarity for places than for people ($\beta_{Category_place}$ = 0.02, *SE* = 0.004, χ^2 = 31.49, p < .001). There was neither a main effect of centrality ($\beta_{Centrality}$ = -0.007, *SE* = 0.004, χ^2 = 0.9, p = .342) nor an interaction of category with centrality ($\beta_{Category_place:Centrality}$ = 0.007, *SE* = 0.004, χ^2 = 3.04, p = .081). The same pattern (i.e., only a main effect of category) also emerged for the model based on affective value (see Supplementary Table 4).

For the PCC, the model based on affective value performed decisively better than any other model: principal component (*LER* = 4.14), experience (*LER* = 6.82), and centrality (*LER* = 33.99) (Fig. 2 d). The main effect of affective value was significant, indicating overall greater neural pattern similarity for nodes of higher affective value ($\beta_{AffectiveValue}$ = 0.026, *SE* = 0.007, χ^2 = 10.71, p = .001). There was no main effect of category ($\beta_{Category_place}$ = 0.014, *SE* = 0.009, χ^2 = 1.51, p = .219), but an interaction of affective value with category, reflecting a stronger effect of affective value for people than places ($\beta_{Category_place:AffectiveValue}$ = -0.01, *SE* = 0.005, χ^2 = 3.92, p = .048) (see Supplementary Table 4 for all model parameters).

In summary, the model based on the principal component was the clear winner in the mPFC, whereas it was outperformed by alternative models in the control regions. This pattern thus suggests some regional specificity of value-weighted schemas. Note that we obtained consistent results when examining the structure of representations in the purely anatomically defined ROIs (see Supplementary Fig. 2; Supplementary Table 4).

Finally, we sought to ensure that the winning models in each ROI perform better null models based on noise. To this end, for each familiar person and place, we randomly sampled a value from a standard normal distribution. We then used these values to construct a noise model by performing the identical processing steps as for the other predictors.

Moreover, we derived a second noise model by sorting the noise vector in descending order prior to constructing the model. This was done to account for the order of columns in the neural RSM that was not randomly determined. Instead, the order reflected the serial position of the people and places as provided by the participants. Participants tend to start listing people and places with the ones that they like and know better. As a consequence, nodes that are listed earlier tend to also have higher values on the principal component. By sorting the noise vectors in descending order, we imposed a similar dependence between the noise values and their serial positions.

Separately for each ROI, we then compared model performance of random noise and sorted noise against the winning model. Because there was no clear single winning model for the hippocampus, we included the models based on cen-

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trality and on affective value (*LER* < 2) in this comparison. We repeated this process 1,000 times to obtain an estimate of the expected performance of the noise models and the winning model. As expected, in the mPFC, the principal component remained the best model (*mean LER* = 0, *SD* = 0.04), performing decisively better than sorted noise (*mean LER* = 15.56, *SD* = 2.51) and random noise (*mean LER* = 25.81, *SD* = 6.03).

For the hippocampus, the initial model comparisons had provided only minimal evidence for centrality over affective value. We therefore compared both of these models with the noise models. Again, there was minimal evidence for superiority of centrality (*mean LER* = 0.09, *SD* = 0.53) over affective value (*mean LER* = 0.58, *SD* = 0.53). Both models performed decisively better than sorted noise (*mean LER* = 4.4, *SD* = 1.09) and random noise (*mean LER* = 5.32, *SD* = 2.38). The model comparisons in the PCC revealed strong, though not decisive, evidence for a superiority of affective value (*mean LER* = 1.15, *SD* = 2.1) over sorted noise (*mean LER* = 2.89, *SD* = 3.37). However, both did perform decisively better than random noise (*mean LER* = 29.46, *SD* = 9.01) (see Fig. 2 e).

The results thus provide further evidence for the hypothesis that the mPFC encodes both the nodes and the edges of value-weighted schematic representations. The model comparison moreover supports this account with some regional specificity.

Discussion

Human adaptive cognition is fostered by representations of the structure of our environment (7, 15, 17). Such structured representations act as templates that allow us to facilitate recollections of the past, to make sense of the present, and to flexibly anticipate the future (6, 8, 13, 33, 62). Structured representations have been described in the mPFC for various domains, ranging from spatial and conceptual to abstract state spaces (63–66).

Our results support the hypothesis that the mPFC supports a specific form of such structured representations: value-weighted schemas of our environment. Generally, the mPFC has long been argued to mediate memory schemas (26, 27, 62), yet the exact contribution of this region has remained unclear. It has been suggested that the mPFC serves to detect congruency of incoming information with schematic knowledge that is represented in posterior areas (27). This region would thus not necessarily represent any kind of schematic knowledge by itself. Our data indicate that the contribution of the mPFC goes beyond congruency detection: It directly encodes schematic representations of the environment (11, see also 19).

These representations could act as pointer functions that guide the reinstatement of relevant distributed information (26, 27, 67, 68). This suggestion fits with broader accounts that situate the mPFC on top of a cortical hierarchy as a convergence zone (48, 69) that integrates information from diverse brain networks (33, 55).

Critically, our results support the hypothesis that schematic representations in the mPFC (e.g., of one's social network) inherently entail the value of the encoded nodes (e.g., how much we like individual people). That is, the structure of the edges could best be accounted for by a model based on a latent factor that quantifies the importance of the nodes. As predicted, this factor was not only influenced by the nodes' centrality (11, 40) and familiarity (19, 29, 33), but also by their affective value (19–21). We obtained this pattern across schemas for personally familiar people and places. The convergent results thus demonstrate that this coding scheme in the mPFC generalizes to different kinds of environmental representations.

The model comparison also suggested some degree of regional specificity for value-weighted schemas. The importance model was neither the best fit to the data obtained from the PCC nor from the hippocampus. Whereas even the best model did not decisively outperform a noise model in the PCC, there was some evidence that the structure of the edges in the hippocampus could best be accounted for by either centrality or affective value. These results are consistent with evidence showing that the hippocampus encodes map-like representations of relational abstract (45, 70) and social (71) knowledge and that it is involved in value learning (72, 73).

More broadly, a functional dissociation between the hippocampus and mPFC is also consistent with the suggested involvement of these regions in two complementary learning systems. Whereas the hippocampus is critical for the retention of individual episodes, the mPFC may extract commonalities across similar events and bind these into consolidated representations (13, 17, 41, 74, cf. 75). The mPFC would thus reduce the complexity of our experience into schematic summary representations.

Indeed, a recent study provided convergent evidence for such dimension reduction in this region. It demonstrated that the mPFC compresses rich perceptual input to only those features that are currently task-relevant – akin to a principal component analysis (14). While such dimension reduction entails the loss of specific details, it also affords generalizability and cognitive flexibility (8, 76). These representations can thus augment planning (6, 77, 78) and also be flexibly used for the construction and valuation of novel events (1, 33).

The emergence of schemas in the mPFC could be fostered by hippocampal replay of past events (79–81). Such replay, conveyed by monosynaptic efferent projections into the mPFC (82), can potentially provide a teaching signal that facilitates neocortical consolidation (41). Moreover, to the degree that replay is biased towards valuable information, it may lead to a stronger weighting of those experiences that are of particular importance (41, 42). However, the mPFC likewise receives

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

direct projections from areas such as the amygdala and the striatum (83) that could also contribute to a shaping of the schematic representations by value (see also 21).

Importantly, the highlighted structure of representations in the mPFC provides a common account for the involvement of this region in both memory schemas and valuation. That is, when we think about an individual element from our environment (e.g., a known person), its representation in the mPFC is activated. This activation then spreads throughout the network of connected nodes. Critically, we suggest that there is a wider spread from nodes that are more valuable and that are thus more strongly embedded in their overarching schema. This wider spread, in turn, may manifest as greater regional univariate activity. According to this account, the valuation signal that has been attributed to the mPFC (30, 34) thus constitutes an emergent property of the structure of its encoded representations.

This interpretation similarly accounts for the stronger engagement of the mPFC when individuals think about themselves as compared to others (84, 85). The self can be considered a super-ordinate schema that entails abstracted representations of all our personal experiences (67). Instantiating this schema would thus presumably lead to wide spread activity, whereas thinking about specific other people would only co-activate neural representations of more restricted nodes. Moreover, the net activity would be lower for other people that we feel less connected to and that we have less experience with (33, 86–88).

To conclude, this study provides evidence that the medial prefrontal cortex represents the structure of our environment in the form of value-weighted schemas. These schemas reflect our experience with individual nodes as well as their centrality. Critically, they also inherently encode information about their affective value. These schematic representations thus prioritize information that is critical for adaptive planning and ultimately promotes our well-being and survival.

Methods

Participants. We recruited 39 right handed healthy unmedicated adults (23 females; *mean* age = 25.4 years, *SD* = 2.6 years) from the study database of the Max Planck Institute for Human Cognitive and Brain Sciences. All participants had normal or corrected to normal vision, provided written informed consent and received monetary compensation for their participation. The experimental protocol was approved by the local ethics committee (Ethical Committee at the Medical Faculty, Leipzig University, Leipzig, Germany; Proposal number: 310/16-ek). Three participants had to be excluded from analysis either because of a recording error ($n = 1$), or excessive movement ($n = 2$). Excessive movement was defined as absolute movement ≤ 3 mm within either run or a total of ≤ 5 episodes of movement ≤ 0.5 mm. We thus included 36 participants (22 females; *mean* age = 25.2 years, *SD* = 2.5 years) in the analyses.

Task and procedures. The procedure, adapted from ref. 19, comprised two sessions. During the first session, participants provided names of personally familiar people and of such places. Participants tend to start by listing people and places that they are most familiar with and that they like the most. We therefore asked them to provide us with 90 people and 90 places and then randomly sampled 30 of each to ensure a greater variability in these variables of interest.

Assessing centrality, experience, and affective value. To quantify the centrality of each node to its schema, participants arranged the names of the people and places on separate two-dimensional circular arenas using the multiple arrangements task (43) (Fig. 1 d). We instructed participants to position names closer to each other that they also associate more strongly. The inverse of the distance thus serves as a measure of associatedness between any two nodes. We quantified the centrality of each person and place to its schema by computing their centrality, i.e., the sum of their associatedness values.

We then assessed how much experience participants had with each person and place. The participants therefore placed the names on continuous familiarity scales ranging from "not at all familiar" to "very much familiar" (Fig. 1 d). Finally, participants provided a measure of affective value for each person and place by arranging their names on continuous liking scales ranging from "not at all liked" to "very much liked". All arrangements were done separately for people and places.

Assessing neural representations. The participants returned for a separate session (*median* delay: 1 day; *range*: 1–4 days) to complete the episodic simulation task in the fMRI scanner. Each trial of the simulation task began with a fixation period of 0.5 s followed by the name of a person or a place for 7.5 s. During this time, participants imagined interacting with the person in a typical manner or being at the place engaging in a location specific activity. Participants were instructed to imagine the episode as vividly as possible, so that they have a clear mental picture of the respective person or place. Participants then rated the vividness of their imagination on a five-point scale within a maximum of 3 s. Trials for which participants failed to press a button within that time period were later removed from analysis. If there was time left from the response window, it was added to the subsequent inter trial interval. This lasted for at least 3 s plus an additional jittered period (0 to 8 s in 2 s intervals). The screen during the inter trial interval was blank. Each person and place was presented once in each of the two functional runs that followed different random orders. Before entering the scanner, participants practiced the simulation task with people and places that they had previously provided but that did not feature in the simulation task proper.

After the simulation task, participants were presented with people-places and faces-places localizers. Outside the scanner, they provided further information regarding the associ-

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ations and identities of the individual people and places, including their addresses and locations. They also completed a number of standard questionnaires. These data were not analyzed for the current study.

fMRI data acquisition. Participants were scanned with a 3 Tesla Siemens Magnetom PRISMA MRI scanner with a 32-channel head coil. We acquired anatomical images with a T1-weighted magnetization-prepared rapid gradient-echo sequence (MPRAGE, 256 sagittal slices, TR = 2,300 ms, TE = 2.98 ms, flip angle = 9°, 1 x 1 x 1 mm³ voxels, FoV = 240 mm, GRAPPA factor = 2). For each of the two functional runs of the simulation task, we acquired 469 volumes of blood-oxygen-level-dependent (BOLD) data with a T2*-weighted echo-planar imaging (EPI) pulse sequence (89, 90). This sequence employed multiband RF pulses with the following parameters: 72 interleaved axial-oblique slices (angled 15° towards coronal from AC-PC), TR = 2,000 ms, TE = 25 ms, flip angle = 90°, 2 x 2 x 2 mm³ voxels, 6/8 partial Fourier, FoV = 192 mm, MF = 3). The first five volumes of each run were discarded to allow for T1 equilibration effects.

fMRI data analysis. Data were analyzed using SPM12 (www.fil.ion.ucl.ac.uk/spm) in Matlab (version 9.3). The functional images were corrected for slice acquisition times, realigned, corrected for field distortions, and co-registered with the anatomical scan. Correction for field distortions was achieved using FSL topup (91, 92) as implemented in FSL 5.0 (https://fsl.fmrib.ox.ac.uk/).

We then decomposed the variance in the BOLD time-series using a general linear model (GLM) in SPM12 (93). Each model included six regressors representing residual movement artifacts, plus regressors modeling the intercepts of block and session. The additional regressors in the GLM coded for the effects of interest.

Specifically, we modeled each trial as a separate condition yielding a total of 120 regressors – one for each of the two simulations of the 30 people and 30 places. The trial regressors were convolved with the canonical hemodynamic response function. A 1/128-Hz high-pass filter was applied to the data and the model. We computed *t*-maps for the estimated parameters of interest (i.e., for each simulation) against the implicit baseline. The ensuing parameters were used for representational similarity analysis (RSA) (44, 59).

Identifying regions that code for nodes of the schemas. To identify brain regions that encode representations of individual people and places (i.e., the nodes of the schemas), we employed an RSA searchlight analysis (spheres with a radius of 8mm, 4 voxels) across all gray matter voxels. This analysis was based on the RSA toolbox 59 and compared activity patterns across functional runs (54, 94). It identified regions where two simulations of the same person or place yielded more similar activity patterns (same-item similarity) than any two simulations of different people or places (different-item similarity). Specifically, we assessed same-item similarity as the Pearson correlation between the activity pattern of the initial simulation of any given node in the first and its repeated simulation in the second run. Different-item similarity was computed as the average correlation of the initial simulation of a node in the first run with all other nodes of the same category (people or places) in the second run. By constraining the different-item similarity to items of the same category, we ensure that it is not affected by general differences in the neural representation of people versus places. Finally, we determined the magnitude of the node coding as the difference score between same- and different-item similarity (19, 54).

This searchlight analysis yielded a node-coding map for each individual participant. For second level analyses, we *Fisher-z*-transformed these maps, normalized them into MNI space using the DARTEL (95) estimated deformation fields, and smoothed them with a Gaussian Kernel of 8 mm radius at full-width-half-maximum. We then masked the smoothed map with the normalized gray matter masks and tested the significance of the node-coding effect using a simple *t*-contrast at each voxel. We used voxel-level inference at $p < 0.05$ (family-wise-error-corrected) and regarded only clusters that comprised at least 30 contiguous voxels.

Examining the structure of the edges. The second RSA examined whether regions that code for the nodes of the schema also code for the predicted relationships between the nodes (i.e., their edges). This analysis thus examined data from regions-of-interest based on the thresholded node-coding map. Note that the two sets of analyses are based on different parts of the neural RSM and on comparisons of model RSMs that are independent from the node-coding model.

For the mPFC, we joined the two rostral and ventral clusters. For the PCC, we took the conjunction of a broad cluster that included this region and an anatomical PCC mask from the Brainnetome atlas (58) (regions 175, 176, 181, 182). For the hippocampal ROI, we merged its rostral and caudal parts of the same atlas (regions 215 – 218). Voxels were included if they had at least 50% probability of being part of the mask and gray matter.

As complementary analyses, we also examined the data solely based on an anatomical mask of the ventral mPFC used by ref. 19 (see also 96), a more spatially extended mask including the rostral mPFC (comprising Brainnetome regions 13, 14, 41, 42, 47 – 50, 187, 188), and said anatomical mask of the PCC. All masks were inverse normalized into subject space using the DARTEL estimated deformation fields and constrained using the implicit mask estimated from the first level GLMs.

Extraction of the importance weights using Principal Component Analysis. We had hypothesized that centrality, experience, and affective value would jointly contribute to the importance of a node and expected that they would share a common latent factor. We thus applied principal component analysis (PCA) to the three features and com-

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

puted the latent factor that explained the most variance. The PCAs were conducted separately for people and places and were based on values of each variable that had been z-scored for each participant. This approach ascertained that neither between-category variance nor between-participant variance would bias the factor solution. We then extracted, across all participants, the respective first principal component for people and places. These principal components were positively correlated not only with centrality and experience but also with affective value, consistent with our proposal that all three contributing features jointly quantify the importance of a given node. We thus refer to these principal components as importance factors.

Predicting the structure of value-weighted schematic representations. We used the importance values to predict the structure of schematic representations in the mPFC. We had hypothesized that more important nodes should, overall, exhibit greater neural similarity with the other nodes. We thus predicted the similarity for any pair of nodes by the product of their respective importance values. We scaled the vectors to the interval of zero (lowest importance) and one (highest importance) prior to multiplication. We then arranged the combined importance values in square matrices for each category (people, places). In initial analyses, we examined the structure of representations separately for the schemas comprising people and places. Note that all analyses are only based on the lower triangular vector of the representational similarity matrices.

Testing the structure of the neural representations using Linear Mixed Effects Models. We set up a series of linear mixed effects models in R (version 3.5.1, www.r-project.org), using LME4 (97), to test which of several alternative predictors accounted best for the structure of representations in the ROIs. These models accounted for the neural similarity data as a function of the full fixed effects of category (people, places) and predictor of interest (i.e., centrality, experience, affective value, or the principal component). We further accounted for between participant variance by including random effects: one random intercept for participant and run as well as random slopes for category and the respective predictor. Hence, all models were of the form:

```
Neural similarity ~ category * predictor +
    (1 + category + predictor | participant) +
    (1 | participant:run)
```

We estimated the models separately for each ROI and subsequently performed model comparisons based on the relative Log Evidence Ratios (*LER*) derived from Akaike's Information Criterion (60). The best model assumes, by definition, a relative *LER* of zero, and we regard relative *LER* differences greater than two as *decisive* evidence for the better model (61).

We further examined whether the winning models in each ROI are also substantially superior to models based on random Gaussian noise. We thus created null models by randomly sampling 30 values from a standard normal distribution for both people and places. We then rescaled these values to the interval from zero to one. Subsequently, we constructed a noise null model by computing the product of every combination of two values, just as we had done for our predictors of interest. We also created a second null model by sorting the same random noise values in descending order prior to multiplication. This was done to account for the inherent order of the original lists of people and places provided by the participants that tended to start with more familiar and pleasant exemplars. Thus, people and places that were named first always received larger random numbers than those named later.

We then fit linear mixed effect models for these two noise null models, and performed a model comparison with the winning model(s) from the respective ROI. We repeated this estimation process 1,000 times to compute average model performance. Critically, if the winning model(s) in each ROI constitute(s) a good approximation of the structure of neural representations, they should consistently outperform both the random noise and the sorted noise models.

ACKNOWLEDGEMENTS

This work was supported by a Max Planck Research Group (awarded to RGB). We thank Roxanne Eisenbeis for assistance in data collection, Ruud Berkers for help in setting up preprocessing, Davide Stramaccia for discussing the linear mixed effects models, and Angharad Williams for comments on an earlier draft of this manuscript.

AUTHOR CONTRIBUTIONS

PCP and RGB designed the study with input from IC. PCP collected the data and wrote the analysis code in consultation with IC and RGB. PCP and RGB analyzed the data and wrote the paper incorporating feedback from IC.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY

Participants did not give consent for their MRI data to be released publicly within the General Data Protection Regulation 2016/679 of the EU. We can thus only share data with individual researchers upon reasonable request. The second level t-map for the node coding searchlight analysis (Fig. 1 c) and the functional masks of the mPFC and PCC ROI are available at neurovault: https://identifiers.org/neurovault.collection:8129

CODE AVAILABILITY

Custom code is publicly available via the Open Science Framework: https://osf.io/6h58e/?view_only=b21fa36180a845e69671f222a110bac8

References

1. Barron, H. C., Dolan, R. J. & Behrens, T. E. J. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* **16**, 1492–1498 (2013).
2. Hassabis, D. & Maguire, E. A. Deconstructing episodic memory with construction. *Trends Cogn. Sci.* **11**, 299–306 (2007).
3. Irish, M., Addis, D. R., Hodges, J. R. & Piguet, O. Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. *Brain* **135**, 2178–2191 (2012).
4. Schacter, D. L., Benoit, R. G. & Szpunar, K. K. Episodic future thinking: mechanisms and functions. *Curr. Opin. Behav. Sci.* **17**, 41–50 (2017).
5. Suddendorf, T. & Corballis, M. C. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behav. Brain. Sci.* **30**, 299–313; discussion 313–351 (2007).

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Addis, D. R. Mental time travel? A neurocognitive model of event simulation. *Rev. Phil. Psych.* **11**, 233–259 (2020).
7. Bartlett, F. C. *Remembering: A study in experimental and social psychology* (Cambridge University Press, Cambridge, 1932).
8. Ghosh, V. E. & Gilboa, A. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia* **53**, 104–114 (2014).
9. Piaget, J. *The origins of intelligence in children* (W.W. Norton & Co, New York, 1952).
10. Chen, C., Lu, Q., Beukers, A., Baldassano, C. & Norman, K. A. Learning to perform role-filler binding with schematic knowledge. *arXiv:1902.09006 [cs]* (2020). URL http://arxiv.org/abs/1902.09006.
11. Parkinson, C., Kleinbaum, A. M. & Wheatley, T. Spontaneous neural encoding of social network position. *Nat. Hum. Behav.* **1**, 0072 (2017).
12. Rumelhart, D. & Ortony, A. The representation of knowledge in memory. In Anderson, R., Spiro, R. & Montague, W. (eds.) *Schooling and the acquisition of knowledge*, 99–135 (Erlbaum, Hilsdale, 1976).
13. Moscovitch, M., Cabeza, R., Winocur, G. & Nadel, L. Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annu. Rev. Psychol.* **67**, 105–134 (2016).
14. Mack, M. L., Preston, A. R. & Love, B. C. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* **11**, 1–11 (2020).
15. Behrens, T. E. J. *et al.* What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
16. Bulley, A. & Schacter, D. L. Deliberating trade-offs with the future. *Nat. Hum. Behav.* **4**, 238–247 (2020).
17. Morton, N. W., Sherrill, K. R. & Preston, A. R. Memory integration constructs maps of space, time, and concepts. *Curr. Opin. Behav. Sci.* **17**, 161–168 (2017).
18. O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, reward, and decision making. *Annu. Rev. Psychol.* **68**, 73–100 (2017).
19. Benoit, R. G., Paulus, P. C. & Schacter, D. L. Forming attitudes via neural activity supporting affective episodic simulations. *Nat. Commun.* **10**, 2215 (2019).
20. Farovik, A. *et al.* Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. *J. Neurosci.* **35**, 8333–8344 (2015).
21. Roy, M., Shohamy, D. & Wager, T. D. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn. Sci.* **16**, 147–156 (2012).
22. Zhou, J. *et al.* Rat orbitofrontal ensemble activity contains multiplexed but dissociable representations of value and task structure in an odor sequence task. *Curr. Biol.* **29**, 897–907.e3 (2019).
23. Brod, G. & Shing, Y. L. Specifying the role of the ventromedial prefrontal cortex in memory formation. *Neuropsychologia* **111**, 8–15 (2018).
24. Ghosh, V. E., Moscovitch, M., Colella, B. M. & Gilboa, A. Schema representation in patients with ventromedial PFC lesions. *J. Neurosci.* **34**, 12057–12070 (2014).
25. Tse, D. *et al.* Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
26. Gilboa, A. & Marlatte, H. Neurobiology of schemas and schema-mediated memory. *Trends Cogn. Sci.* **21**, 618–631 (2017).
27. van Kesteren, M. T. R., Ruiter, D. J., Fernández, G. & Henson, R. N. How schema and novelty augment memory formation. *Trends Neurosci.* **35**, 211–219 (2012).
28. Thornton, M. A. & Mitchell, J. P. Theories of person perception predict patterns of neural activity during mentalizing. *Cereb. Cortex* **28**, 3505–3520 (2018).
29. Robin, J., Buchsbaum, B. R. & Moscovitch, M. The primacy of spatial context in the neural representation of events. *J. Neurosci.* **38**, 2755–2765 (2018).
30. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
31. Peters, J. & Büchel, C. Neural representations of subjective reward value. *Behav. Brain. Res.* **213**, 135–141 (2010).
32. Winecoff, A. *et al.* Ventromedial prefrontal cortex encodes emotional value. *J. Neurosci.* **33**, 11032–11039 (2013).
33. Benoit, R. G., Szpunar, K. K. & Schacter, D. L. Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proc. Natl. Acad. Sci. USA* **111**, 16550–16555 (2014).
34. Clithero, J. A. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014).
35. Lin, W.-J., Horner, A. J., Bisby, J. A. & Burgess, N. Medial prefrontal cortex: adding value to imagined scenarios. *J. Cogn. Neurosci.* **27**, 1957–1967 (2015).
36. Camille, N., Griffiths, C. A., Vo, K., Fellows, L. K. & Kable, J. W. Ventromedial frontal lobe damage disrupts value maximization in humans. *J. Neurosci.* **31**, 7527–7532 (2011).
37. Fellows, L. K. The functions of the frontal lobes: evidence from patients with focal brain damage. *Handb. Clin. Neurol.* **163**, 19–34 (2019).
38. Lim, S.-L., O'Doherty, J. P. & Rangel, A. Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. *J. Neurosci.* **33**, 8729–8741 (2013).
39. Shenhav, A., Barrett, L. F. & Bar, M. Affective value and associative processing share a cortical substrate. *Cogn. Affect. Behav. Neurosci.* **13**, 46–59 (2013).
40. Zerubavel, N., Bearman, P. S., Weber, J. & Ochsner, K. N. Neural mechanisms tracking popularity in real-world social networks. *Proc. Natl. Acad. Sci. USA* **112**, 15072–15077 (2015).
41. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
42. Schafer, M. & Schiller, D. Navigating social space. *Neuron* **100**, 476–489 (2018).
43. Kriegeskorte, N. & Mur, M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 245 (2012).
44. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2** (2008).
45. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* **6**, e17086 (2017).
46. Josselyn, S. A. & Frankland, P. W. Memory allocation: mechanisms and function. *Annu. Rev. Neurosci.* **41**, 389–413 (2018).
47. Sawamura, H., Orban, G. A. & Vogels, R. Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the fMRI adaptation paradigm. *Neuron* **49**, 307–318 (2006).
48. Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R. & Buckner, R. L. Functional-anatomic fractionation of the brain's default network. *Neuron* **65**, 550–562 (2010).
49. Baldassano, C., Hasson, U. & Norman, K. A. Representation of real-world event schemas during narrative perception. *J. Neurosci.* **38**, 9689–9699 (2018).
50. Benoit, R. G. & Schacter, D. L. Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia* **75**, 450–457 (2015).
51. Maguire, E. A., Frith, C. D. & Morris, R. G. M. The functional neuroanatomy of comprehension and memory: the importance of prior knowledge. *Brain* **122**, 1839–1850 (1999).
52. Grueschow, M., Polania, R., Hare, T. A. & Ruff, C. C. Automatic versus choice-dependent value representations in the human brain. *Neuron* **85**, 874–885 (2015).
53. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. USA* **111**, 14565–14570 (2014).
54. Nili, H., Walther, A., Alink, A. & Kriegeskorte, N. Inferring exemplar discriminability in brain representations. *PLoS ONE* **15**, e0232551 (2020).
55. Ritchey, M., Libby, L. A. & Ranganath, C. Cortico-hippocampal systems involved in memory and cognition: the PMAT framework. *Prog. Brain. Res.* **219**, 45–64 (2015).
56. Rugg, M. D. & Vilberg, K. L. Brain networks underlying episodic memory retrieval. *Curr. Opin. Neurobiol.* **23**, 255–260 (2013).
57. Milivojevic, B., Vicente-Grabovetsky, A. & Doeller, C. F. Insight reconfigures hippocampal-prefrontal memories. *Curr. Biol.* **25**, 821–830 (2015).
58. Fan, L. *et al.* The Human Brainnetome Atlas: a new brain atlas based on connectional architecture. *Cereb. Cortex* **26**, 3508–3526 (2016).
59. Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
60. Snipes, M. & Taylor, D. C. Model selection and Akaike Information Criteria: an example from wine ratings and prices. *J. Wine. Econ.* **3**, 3–9 (2014).
61. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
62. Burgess, P. W. & Shallice, T. Confabulation and the control of recollection. *Memory* **4**, 359–411 (1996).
63. Constantinescu, A. O., OReilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
64. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
65. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nat. Neurosci.* **16**, 486–492 (2013).
66. Schuck, N., Cai, M., Wilson, R. & Niv, Y. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).
67. Gilboa, A. & Moscovitch, M. Ventromedial prefrontal cortex generates pre-stimulus theta coherence desynchronization: a schema instantiation hypothesis. *Cortex* **87**, 16–30 (2017).
68. Ciaramelli, E., De Luca, F., Monk, A. M., McCormick, C. & Maguire, E. A. What "wins" in VMPFC: Scenes, situations, or schema? *Neurosci. Biobehav. Rev.* **100**, 208–210 (2019).
69. Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. USA* **113**, 12574–12579 (2016).
70. Theves, S., Fernandez, G. & Doeller, C. F. The hippocampus encodes distances in multidimensional feature space. *Curr. Biol.* **29**, 1226–1231.e3 (2019).
71. Tavares, R. M. *et al.* A map for social navigation in the human brain. *Neuron* **87**, 231–243 (2015).
72. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
73. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
74. Norman, K. A. & O'Reilly, R. C. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol. Rev.* **110**, 611–646 (2003).
75. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B.* **372**, 20160049 (2017).
76. Bowman, C. R. & Zeithamova, D. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *J. Neurosci.* **38**, 2605–2614 (2018).
77. Momennejad, I. Learning structures: predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* **32**, 155–166 (2020).
78. Günseli, E. & Aly, M. Preparation for upcoming attentional states in the hippocampus and medial prefrontal cortex. *eLife* **9**, e53191 (2020).
79. Carr, M. F., Jadhav, S. P. & Frank, L. M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* **14**, 147–153 (2011).
80. Michelmann, S., Staresina, B. P., Bowman, H. & Hanslmayr, S. Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nat. Hum. Behav.* **3**, 143–154 (2019).
81. Wimmer, G. E., Liu, Y., Vehar, N., Behrens, T. E. J. & Dolan, R. J. Episodic memory retrieval success is associated with rapid replay of episode content. *Nat. Neurosci.* **23**, 1025–1033 (2020).
82. Eichenbaum, H. Prefrontal–hippocampal interactions in episodic memory. *Nat. Rev. Neurosci.* **18**, 547–558 (2017).
83. Price, J. L. & Drevets, W. C. Neurocircuitry of mood disorders. *Neuropsychopharmacol.* **35**, 192–216 (2010).
84. Northoff, G. *et al.* Self-referential processing in our brain - A meta-analysis of imaging

PREPRINT

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.21.260950; this version posted December 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

studies on the self. *Neuroimage* **31**, 440–457 (2006).
85. Overwalle, F. V. Social cognition and the brain: a meta-analysis. *Hum. Brain. Mapp.* **30**, 829–858 (2009).
86. Benoit, R. G., Gilbert, S. J., Volle, E. & Burgess, P. W. When I think about me and simulate you: medial rostral prefrontal cortex and self-referential processes. *Neuroimage* **50**, 1340–1349 (2010).
87. Mitchell, J. P., Macrae, C. N. & Banaji, M. R. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* **50**, 655–663 (2006).
88. Rameson, L. T., Satpute, A. B. & Lieberman, M. D. The neural correlates of implicit and explicit self-relevant processing. *Neuroimage* **50**, 701–708 (2010).
89. Feinberg, D. A. *et al.* Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PLoS ONE* **5**, e15710 (2010).
90. Moeller, S. *et al.* Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med.* **63**, 1144–1153 (2010).
91. Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* **20**, 870–888 (2003).
92. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, 208–219 (2004).
93. Friston, K. J. *et al.* Analysis of fMRI time-series revisited. *Neuroimage* **2**, 45–53 (1995).
94. Alink, A., Walther, A., Krugliak, A., Bosch, J. J. F. v. d. & Kriegeskorte, N. Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *bioRxiv* 032391 (2015). URL https://www.biorxiv.org/content/10.1101/032391v2.
95. Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95–113 (2007).
96. Liu, Z.-X., Grady, C. & Moscovitch, M. Effects of prior-knowledge on brain activation and connectivity during associative memory encoding. *Cereb. Cortex* **27**, 1991–2009 (2017).
97. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* **67** (2015).

PREPRINT