Neural representation of abstract task structure during generalization

Authors: Avinash R. Vaidya [1], Henry M. Jones [1,2], Johanny Castillo [1,3], & David Badre [1,4]

[1] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
[2] Department of Psychology, Stanford University
[3] Department of Psychology and Brain Sciences, University of Massachusetts Amherst
[4] Carney Institute for Brain Science, Brown University

Address correspondence to:
Avinash R. Vaidya
Brown University
Providence, RI 02912-1978
tel. 401-863-3634
email: avinash_vaidya (at) brown.edu

Acknowledgements: This work was supported by a Multidisciplinary University Research Initiative award from the Office of Naval Research (N00014-16-1-2832), a Postbaccalaureate Research Education Grant (R25GM125500), and a Ruth L. Kirschstein National Research Service Award to A.V. (F32MH116592). We thank Apoorva Bhandari and other members of the Badre Lab for constructive discussion and consultation on task design and fMRI analysis, as well as Linda Q. Yu for comments on this manuscript. We also thank Emily Chicklis for assistance with data collection.

Abstract

Abstract task representations enable generalization, including inferring new behaviors based on prior knowledge without additional training. However, evidence for a neural representation that meets this benchmark is surprisingly limited. Here, using functional MRI (fMRI), we observed that abstract task structure was represented within frontoparietal networks during generalization. These results reveal the neural systems supporting a vital feature of human cognition: the abstraction of task knowledge to infer novel behaviors.

Main text

Many complex tasks we perform daily, though different in their details, share an abstract structure. It has long been proposed that humans can learn this abstract structure and leverage it to think creatively, make novel inferences and rapidly generalize knowledge to unique problems we have never encountered before [1–3]. Recent work has provided evidence that the orbitofrontal cortex (OFC) and hippocampus (HPC) maintain a structured, abstract representation of tasks [4–7]. Likewise, other work has investigated the neural processes supporting fast acquisition of stimulus-response rules based on latent task knowledge [8,9]. However, a defining feature of an abstract representation is that it can be used to generalize behaviors to new settings in absence of feedback, through a process of inference. To date, no study has investigated the neural correlates of an abstract task representation that is observed to satisfy this criterion. As a consequence, the neural systems supporting this essential function remain unknown. To address this gap, we used fMRI to test the hypothesis that latent task states are instantiated in neural activity during generalization, particularly in OFC and HPC.

Participants completed a task where they gathered rewards in an environment where a latent, generalizable structure was available. Throughout the experiment, images of trial-unique items from three categories appeared beneath a context, denoted by a scene **(Figure 1a**). Participants decided whether to "sell" each item or pass. If they sold, they would receive or lose reward probabilistically, as determined by the combination of item category and context. Participants saw each of these combinations in short batches of trials termed mini-blocks. Importantly, the contexts can be clustered based on the expected values associated with each item category. This structure provided an opportunity to link these contexts together using an abstract, latent state representation. We hypothesized that participants would form and later use this representation to generalize to new conditions.

This hypothesis was tested across three phases (**Supplementary Figure 1**). During phase one, participants learned about the abstract latent states using three item categories (hands, foods and leaves) across nine contexts (**Figure 1b**). In the next phase, participants learned the values of three new item categories (faces, animals, and objects) in three of the nine contexts; one from each cluster (**Figure 1c**). Then, in a final generalization phase, they were tested on the remaining six contexts using the three new image categories, without feedback. Optimal performance depended on generalizing the new category values learned during the second phase to the held-out contexts (**Figure 1d**). This inference was only possible if those contexts were linked to a latent state representation formed during the first phase.

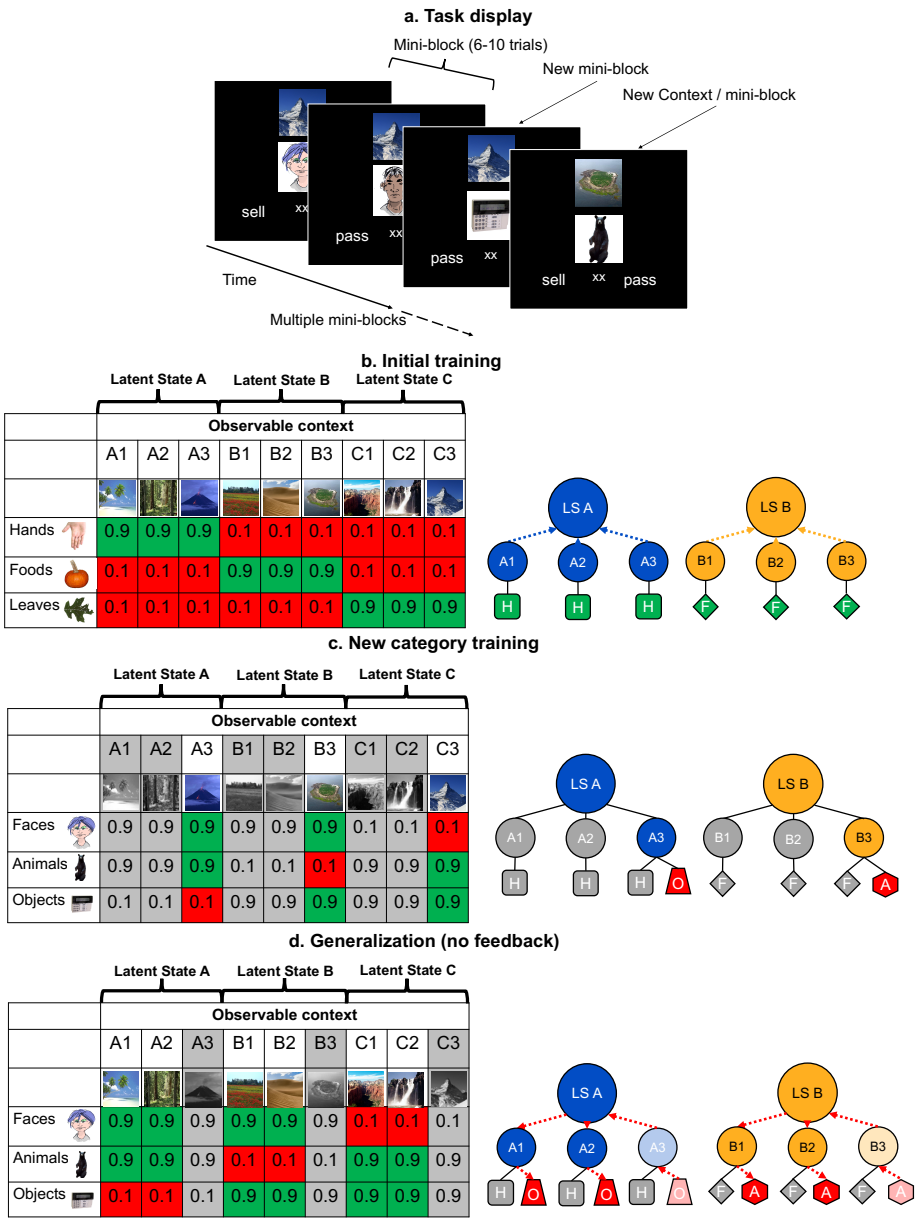

Figure 1. Schematic of the experimental task, and its design and logic. **a.** In each trial of the training and generalization phases, participants were asked to make a decision to sell or pass on an image, the value of which depended on contexts shown above the image. Throughout the two training phases, participants received feedback on every trial, but not during generalization. Participants saw three categories of images in the same context over small batches of trials for each unique combination (termed a mini-block) before switching to a new context. **b-d**. The left tables show the reward structure for example context-category pairs across the three phases of the experiment. Cells show the probabilities of reward for each pair. The right schematics illustrate clustering of contexts by category-values into latent states (blue and orange arrows) and inference of values via structured knowledge (red arrows) with just two latent states. Latent State (LS) Hands (H), Foods (F), Objects (O) and Animals (A). **b.** In the initial training phase, participants were presented trial-unique images from three categories of images. These contexts could be grouped together through an abstract latent state representation based on the similarity of their category-value associations. **c.** Participants were later trained on three new categories in a subset of the previous contexts. Greyed out columns indicate contexts that were left out of this phase. Thus, the values of new categories were trained in only one context. in each LS cluster **d.** In the generalization phase, participants were asked to make decisions about the left out, novel context-category combinations without feedback. Participants had to use their knowledge of the latent states linking contexts together learned in the initial training. Example face images are shown as cartoons rather than the photos of real individuals used in the experiment, in accordance with bioRxiv policy.

Participants were tested in three sessions. In session 1, they performed the task behaviorally. Generalization performance determined their inclusion in the fMRI experiment. 48% of participants passed a criterion of ≥70% accuracy in all generalization conditions and so were recruited for two fMRI sessions. Participants who failed to meet this criterion performed two behavioral sessions, instead. Of these participants, 50% ultimately passed the accuracy criterion during generalization. Thus, the majority of participants (75%) could carry out this generalization task given sufficient experience (**Supplementary Figure 2**).

In session 2, all participants carried out the same task with new context stimuli. For fMRI participants, the generalization phase was completed in the scanner. In session 3, fMRI participants completed a shortened version of the training as a reminder of the task, and continued the generalization phase in the scanner.

Analysis of participants' learning curves demonstrates that participants immediately made use of the latent task structure in the generalization phase. Through the first few trials of each mini-block of the initial and new category training, fMRI participants' accuracy steadily improved. In contrast, in the generalization phase, accuracy was near ceiling from the first trial (mean P(correct) = 0.96, SD = 0.06; **Figure 2a**). The rate of this learning curve in the generalization phase was significantly lower than either of the other phases (**Figure 2b**).

Notably, high accuracy during generalization was accompanied by elevated reaction times (RT; mean = 4.04 s, SD = 1.7 s) on the first trial, with a sharp drop-off subsequently (mean=1.15 s, SD = 0.08 s; **Figure 2c**). Rate of speeding was significantly steeper in the generalization phase (**Figure 2d**). Thus, rather than immediately apply a structure learned incidentally during training (phase 2), participants took time to infer values during the first generalization trials. However, participants were faster for the second presentation of contexts from the same latent state during generalization compared to contexts from different latent states, indicating that additional inferences were made more quickly when a latent state had been accessed previously (**Supplementary Figure 3**).

After an initial inference, participants may have formed context-category rules and no longer used the latent state representation throughout the scanned generalization task. We reasoned that if participants are continuing to utilize the latent task structure, this would be evident in switch costs when that latent structure changed between-trials [10]. Isolating this analysis to trials where the context switched, we found a significant switch cost in RT for trials when the latent state changed versus stayed the same between trials (repeated-measures t-test: $t(15) = 2.33$, $P = 0.03$, $d = 0.58$; **Figure 2e**), which did not differ between scanner sessions (repeated-measures t-test: $P = 0.8$). These data indicate that participants used these latent state representations throughout the scanned task. Participants were also faster when both the context and latent state remained the same ($t(15) = 5.29$, $P < 0.0001$, $d = 1.32$).

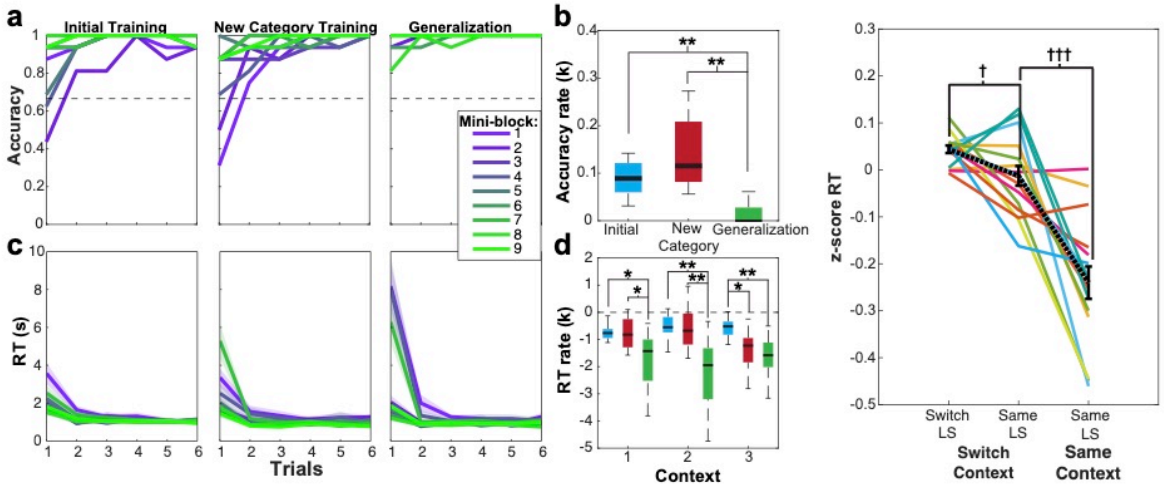

Figure 2. Learning curves and repetition effects from training and generalization phases for fMRI participants in session 2. **a**. Mean accuracy across participants for first six trials of the first nine mini-blocks of initial and new category training, as well as blocked generalization phase. Dashed line indicates chance-level performance. **b.** Boxplots representing the estimated exponential rate of change in mean accuracy in the first six trials for all mini-blocks in each phase, across participants. **c.** Mean reaction times (RT) for each trial in each of the first nine mini-blocks for each of these phases. Shaded area represents standard error of the mean (SEM). Mean and standard error were calculated from the log-transformed RTs. **d.** Boxplots representing estimated exponential rate of change in RT for first six trials within each phase for the first mini-blocks within the first three contexts in presentation order. Boxplots show 25th, 50th and 75th quantiles, whiskers are 1.5 times the interquantile range. * $P < 0.05$, ** $P < 0.01$, Wilcoxon signed rank test, corrected for multiple comparisons. **e.** RTs for trials from the pseudo-randomized generalization phase from sessions 2 and 3 during fMRI scanning where latent states remained the same or switched from the previous trial, while the context switched or stayed the same. RTs have been log-transformed and z-scored within session for each participant. Each line represents a participant, black dashed line indicates mean and error bars indicate SEM. †$P < 0.05$, †††$P < 0.0001$, repeated-measures t-test.

We carried out a representational similarity searchlight analysis (RSA) using multiple linear regression to compare empirical representational dissimilarity matrices (RDMs) from pattern activity to hypothesis RDMs quantifying the predicted distances between conditions based on latent states, contexts, item categories, expected value, interactions between these factors and control regressors (**Supplementary Figure 4**).

We found evidence for a latent state representation in bilateral dorsal- and ventrolateral prefrontal cortex (PFC) with a rostral distribution, precuneus, left middle temporal gyrus and bilateral inferior parietal lobules (**Figure 3a**; **Supplementary Table 1**). Comparing this statistical map with resting-state functional networks from Yeo et al. [11] revealed that this latent state representation overlapped most with a frontoparietal network centered on the inferior frontal and intraparietal sulci (**Supplementary Figure 5**).

In contrast, context was associated with activity in the bilateral fusiform gyri (**Figure 3b**). Expected value was associated with activity in OFC, as well as left superior frontal and angular gyri (**Figure 3***c*). Item category was robustly represented throughout visual areas and PFC (**Supplementary Figure 6**). Interactions between item category with value and latent state were not significant, though items with same value in the same latent state were

represented more similarly in ventral temporal cortex (**Supplementary Figure 7**). These effects were stable across sessions (**Supplementary Table 2**).

We also conducted ROI-based RSAs focused on HPC and OFC, given *a priori* expectation that these regions are involved in representing latent task structure. RSAs within these ROIs were null for both value and latent state representations **(Supplementary Figure 8**). To test these effects at a finer grain, we conducted second-level tests of the whole-brain searchlight analysis masked by these ROIs. Here, we found signals consistent with latent state and expected value representations in central OFC, but not HPC (**Supplementary Figure 9; Supplementary Table 3**).

Lastly, we tested a univariate contrast of correct and error responses. This contrast found robust activation in the HPC and OFC on correct responses (**Supplementary Figure 10**; **Supplementary Table 4**).

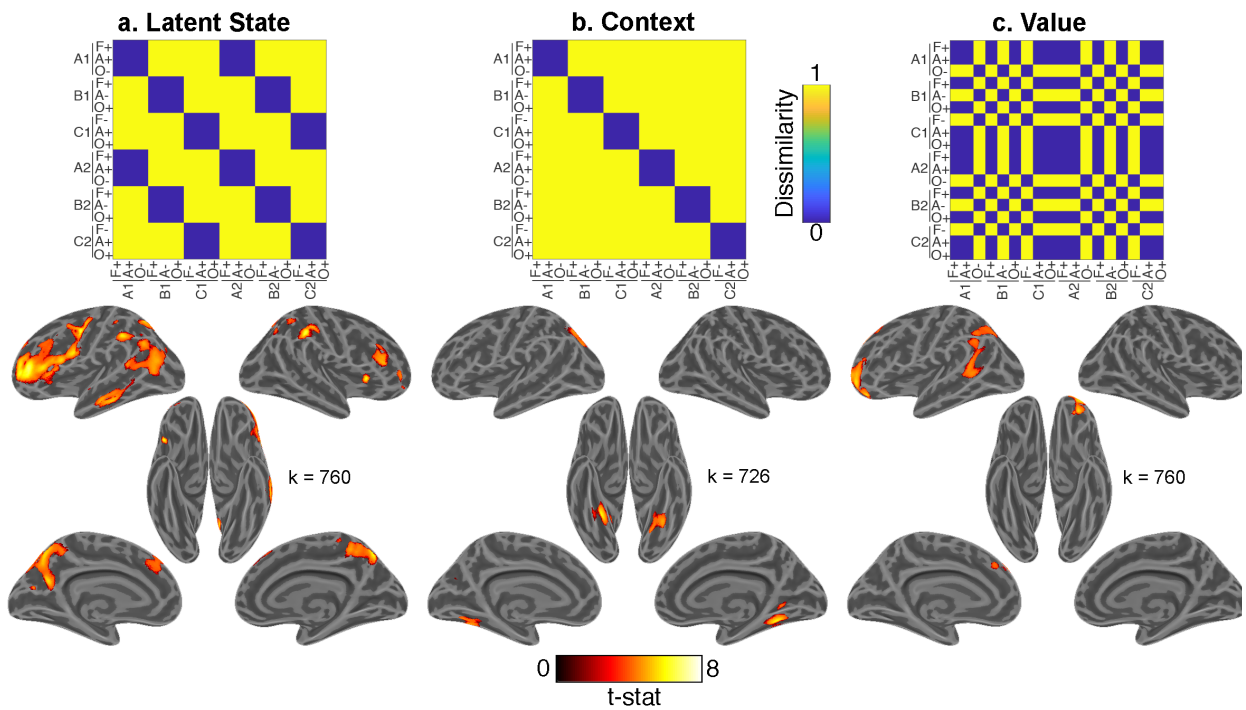

Figure 3. Whole-brain representational similarity searchlight analysis for main effects of interest. Each upper panel shows hypothesis representational dissimilarity matrix (RDM) for task factors. Lower panels show t-statistic map from a searchlight analysis testing these predictions in pattern activity projected onto inflated cortical surfaces. All maps are defined with a cluster forming threshold of $P < 0.001$ and corrected for multiple comparisons with permutation tests for defining a cluster extent threshold at $P < 0.05$. These maps include representations of **a**. latent states **b.** contexts, and **c.** value. Cluster extent threshold for each contrast is given by the value of k. A, B, C refer to distinct latent states. A1-C1 and A2-C2 refer to distinct contexts that belong to each of those latent states. F, Faces; A, Animals; O, Objects. +, positive value; -, negative value.

Here, we observed that a latent state representation capable of supporting generalization of knowledge to new settings is represented within mid-lateral PFC and parietal cortex. This pattern overlaps with a network previously implicated in contextual modulation of behavior in hierarchical reinforcement learning and cognitive control tasks [8,12,13]. Like hierarchical task structure in those experiments, latent task states enable faster

learning, reduced memory load and greater behavioral flexibility [14,15]. Our findings imply that circuitry guiding behavior based on structural knowledge of a task are recruited to facilitate new behaviors during generalization.

We were also able to cross two factors previously associated with OFC function within the same experiment: latent task states and expected value [16]. We found limited evidence for OFC representing latent task states, though expected value was robustly represented in this ROI. These data are consistent with rodent electrophysiological data demonstrating that expected value signals contribute more to the variance of activity within this region than representations of task space [4,17]. However, both OFC and HPC were activated for correct responses compared to errors, possibly consistent with their role in retrieving task knowledge, if not in actively representing latent structure. Our results may be reconciled with prior observations if OFC and HPC are engaged more by initially bridging contexts [18], or resolving uncertainty about the prevailing latent state [19,20].

In sum, we have discovered evidence of a neural instantiation of a long-supposed construct of cognitive models: an abstract task representation that enables generalization to new task conditions. This study opens new ground for understanding how abstract task representations support adaptive inferences.

References

1. Harlow, H. F. The formation of learning sets. *Psychol. Rev.* (1949). doi:10.1037/h0062474
2. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science* (2011). doi:10.1126/science.1192788
3. Tolman, E. C. Cognitive Maps in Rats and Men. *Psychol. Rev.* **55**, 189–208 (1948).
4. Zhou, J. *et al.* Complementary Task Structure Representations in Hippocampus and Orbitofrontal Cortex during an Odor Sequence Task. *Curr. Biol.* (2019). doi:10.1016/j.cub.2019.08.040
5. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
6. Schiller, D. *et al.* Memory and Space: Towards an Understanding of the Cognitive Map. *J Neurosci* **35**, 13904–13911 (2015).
7. McKenzie, S. *et al.* Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron* **83**, 202–215 (2014).
8. Collins, A. G. E., Cavanagh, J. F. & Frank, M. J. Human EEG uncovers latent generalizable rule structure during learning. *J. Neurosci.* (2014). doi:10.1523/JNEUROSCI.3900-13.2014
9. Tomov, M. S., Dorfman, H. M. & Gershman, S. J. Neural computations underlying causal structure learning. *J. Neurosci.* (2018). doi:10.1523/JNEUROSCI.3336-17.2018
10. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* (2013). doi:10.1037/a0030852
11. Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* (2011). doi:10.1152/jn.00338.2011
12. Choi, E. Y., Drayna, G. K. & Badre, D. Evidence for a functional hierarchy of association networks. *J. Cogn. Neurosci.* (2018). doi:10.1162/jocn_a_01229
13. Badre, D., Kayser, A. S. & D'Esposito, M. Frontal cortex and the discovery of abstract action rules. *Neuron* (2010). doi:10.1016/j.neuron.2010.03.025
14. Frank, M. J. & Badre, D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb Cortex* **22**, 509–526 (2012).
15. Koechlin, E. & Summerfield, C. An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci.* (2007). doi:10.1016/j.tics.2007.04.005
16. Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nat Neurosci* **18**, 620–627 (2015).
17. 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).
18. Zeithamova, D., Dominick, A. L. & Preston, A. R. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168–179 (2012).
19. Chan, S. C. Y., Niv, Y. & Norman, K. A. A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex. *J. Neurosci.* **36**, 7817–7828 (2016).
20. Saez, A., Rigotti, M., Ostojic, S., Fusi, S. & Salzman, C. D. Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. *Neuron* **87**, 869–881 (2015).

Methods

Experimental procedure

The study design and analyses were pre-registered on the Open Sciences Framework prior to collecting the data (https://osf.io/x6fmb). Participants completed an acquired equivalence task where they learned that nine contexts belonged to three different latent states (three per state) on the basis of their shared value associations for three object categories. They then used this knowledge to complete a generalization task where they inferred values for a new set of categories in contexts with which they had not been previously paired. A schematic overview of the structure of this experiment is provided in **Figure 1** and **Supplementary Figure 1**.

To aid in building a structured representation of the task, participants were told they were playing the role of a photographer selling photos of different item categories in three goblin kingdoms with distinct preferences for these categories. In each trial, participants had to make a bet on whether to sell an image or to pass. This decision was a risky choice where selling could be paid off with a reward in fictitious gold or punished with a loss of gold. After a decision to sell, feedback was provided in terms of the gold won or lost. Passing would lead to no change in gold, but feedback was provided in terms of the outcome had the participant chosen to sell. Thus, in all phases of the experiment except for generalization (see below), fully informative feedback was provided independent of participants' choices. Participants were incentivized with a monetary bonus proportional to the amount of gold they earned in the task. All photos were trial-unique so that participants never learned the values of specific images, but of general image categories.

This experiment took place over three sessions. Each experimental session was comprised of three main phases. First, in the beginning of the initial training phase, participants were tasked with selling photos of hands, foods and leaves in three contexts, represented by a natural scene image above the photo they were selling (A1, B1, C1). Participants were told that the contexts each came from a different "kingdom", equivalent to the generative latent states (LS A, LS B, LS C), and that each of these kingdoms preferred one category of items strongly over the other two categories. Within each context one item category was associated with a 90% probability of reward and 10% probability of punishment for a decision to sell, while the other two categories were associated with a 10% chance of reward and 90% chance of punishment. In each trial in this initial training phase, participants could gain 50 gold or lose 25 gold for choosing to sell a photo. Thus, the expected value of selling, without knowledge of the probability of reward for each category, was zero.

After two blocks of this training, participants then encountered 'roadwork' after which they would learn about a new context in each kingdom represented by novel natural scenes (A2, B2, C2). This roadwork repeated a second time after another two blocks so that participants learned about three contexts in each of the three kingdoms. Importantly, all contexts associated with a given latent state had the same reward probability structure across item categories. Participants were not directly informed of which contexts corresponded to which kingdoms, but had to make these inferences based on their shared category values. The scene images for each context were randomly selected from a set of nine for each participant, so there was no systematic visual relationship between the contexts of each kingdom. Further, as contexts belonging to each "kingdom" were encountered after

each episode of roadwork, this event bound was not a cue to task structure and no aspect of temporal adjacency could link contexts to a latent state.

These initial training blocks consisted of 54 trials with six trials per condition. This initial training phase was organized in batches consisting of a series of trials for each context-category combination (i.e. a 'mini-block'). These mini-blocks were nested by context, so that participants saw each category in the same context one after another (e.g. A1-hands, A1-leaves, A1-foods), before switching to a new context and looping once again through each combination of context-category pairs (e.g. B1-leaves, etc.). The order of presentation for these categories within each context was randomized, as was the order of these contexts. Following these six initial training blocks, participants completed a final reminder training block with the three categories presented in all nine contexts in mini-blocks of eight trials (216 trials total).

In the second phase, participants completed training on new image categories (faces, animals and man-made objects) within a subset of the previous contexts (e.g. A3, B3, C3). Which set of contexts was presented in this phase (i.e. A1, B1, C1 versus A2, B2, C2 or A3, B3, C3) was randomized across participants. Participants completed 90 trials in the course of a single block, consisting of 10 trial mini-blocks for each combination of the three contexts and three categories.

In the new category training phase, two categories of items were now associated with a 90% chance of reward and 10% chance of punishment in each context, while these values were reversed in one context. The categories that were rewarded and punished in each context were randomized with respect to the categories in the initial training across participants. For example, there was no rule in the experiment that faces have high value in contexts where hands have low value. The punishment associated with selling a photo was increased to 100 gold in this block so that the expected value of selling a photo without knowing the probability of reward in a specific category-context combination was still zero.

Finally, in the third phase, participants carried out a generalization test where they were presented with photos from these new categories in conjunction with the six contexts held out of the prior new category training (e.g. A1, B1, C1 and A2, B2, C2, if A3, B3, C3 were trained with the new categories). Critically, participants did not receive feedback throughout this phase and could thus not learn the values of these new categories from experience. Instead, they would have to rely on their knowledge about which contexts belonged to which kingdoms (i.e. latent states) based on the shared category value associations learned in the initial training phase. This generalization test included 10 trial mini-blocks for each of the 18 conditions (each context-category combination, with three categories and six contexts), resulting in 180 trials total. Participants were informed that the value in gold earned and lost on each trial had been doubled from the new category training phase to further incentivize performance.

It is important to emphasize that generalization during this phase was only possible based on a representation of the shared latent states among contexts. The particular pairings of items and contexts during generalization had never been encountered previously. And, as no feedback was provided during this phase, mappings could not be learned through reinforcement. Further, as the item categories themselves were different from the initial training phase, nothing about the objects or values could link the contexts encountered during generalization to what was learned during phases 1 or 2 except for an abstract latent state. That latent state already clustered the contexts encountered during generalization

with that encountered during new category learning because they had shared a value structure during initial learning, albeit a different value structure on different object categories. Thus, performing accurately during the generalization phase of this task provides unambiguous evidence for reliance on an abstract latent state representation.

For all blocks of the training and mini-blocked generalization phases, there was no response deadline. The inter-trial interval (ITI) was generated from a lognormal distribution with a mean of 1 s, maximum time of 4 s and minimum of 0.5 s. Participants' choice was underlined and the stimuli remained on screen for a 0.5 s inter-stimulus interval. In all blocks, except the generalization test, feedback was displayed for 1.5 s, after which the trial would end. Participants also briefly saw the total gold they had earned within the block after it had ended. However, as there was no way for participants to link these earnings back to the specific trials within the generalization phase, they could not learn the structure of the task from this kind of feedback once they had reached this section of the experiment.

As already noted, this experiment took place over three sessions. The first session was entirely behavioral and served to test how well participants could complete the generalization phase. Prior behavioral pilot experiments had found that approximately 50% of participants did not successfully infer all new context-category values during generalization based on only one experience with training, and so failed to completely learn the latent task structure. As our pre-registered analysis plan depended on participants understanding this structure, we planned*, a priori*, to exclude from fMRI any participants who did not meet a performance criterion of ≥70% accuracy in all 18 conditions of the generalization test in this first session. Instead these participants were asked to return and complete the rest of the experiment in two additional behavioral sessions. This allowed us test whether these participants could generalize in this task, given sufficient experience (see main text). That observation helps limit any concerns about generalizability raised by our selection procedures.

In the second and third session, after completing the blocked generalization test, participants completed three functional runs of the generalization task but without the nested trial structure (i.e. mini-blocks nested by contexts). These blocks also had a three second response deadline for each trial. In each run, ten trials from each of the 18 conditions were presented in a pseudo-random sequence optimized for efficiency using Optseq2 with ITIs ranging from 1 to 9 s (180 trials total) [21]. After each response, the chosen option ('sell' or 'pass') was underlined and the stimulus would remain on-screen until the end of the response deadline. A random subset of six out of 12 optimized sequences was selected for the six functional runs for each participant [21]. Thus, each participant completed 1080 trials of this generalization task over the course of two sessions in the scanner (60 trials per 18 conditions), or behaviorally for those participants who did not pass the generalization criterion in the first session. This dense sampling per participant provided more statistical power for our study by reducing within-subject measurement error. Before completing this task, participants also completed a short practice task with the same mixed trial structure with just one trial from each condition.

Participants were given ample information about the structure of the task in instructions that preceded each new phase of the experiment, but not specific information that about which categories were rewarding in which contexts, and which contexts were part of the same kingdoms (i.e. latent states). Specifically, participants were informed that the outcomes were stable but probabilistic (without being explicitly informed of this

probability), and were told how many categories of photos each kingdom preferred in each phase of the task. Participants were also instructed that they would be 'tested' later without feedback, so they should strive to commit the values of photo categories in each context to memory. They were also informed of the latent task structure in a general sense, namely that contexts represented locations within distinct kingdoms, and that the relation of contexts to different kingdoms could be uncovered by their shared values for the item categories, but were never explicitly told which contexts were associated with which kingdoms. Participants were also informed that the contexts within each kingdom could appear to be visually very different from each other to discourage them from using such a strategy to link contexts together. To facilitate forming such an abstract task representation, the instructions specifically suggested that participants use a semantic elaboration strategy that tied together the distinct contexts within each kingdom, and the values of the categories within these contexts. These instructions were included to help encourage participants to learn the abstract structure of the task rather than try and individually commit the values of all 54 context-category conjunctions to memory.

At the end of the third session, outside of the scanner, participants completed a similarity judgment task for the context natural scene images that were shown in their specific generalization test. Participants were shown these images on a black background in random starting positions and asked to use the mouse to drag and drop these images on-screen so that their distances reflected the similarity. Participants were specifically instructed not to use any information about to which kingdoms the contexts belonged, but only the visual content of the image. The Euclidean distances between these images was then used to estimate a subjective, participant-specific estimate of the visual similarity of these images used as a nuisance regressor for RSA analyses.

Participants

Fifty participants were recruited for this study. Four participants were found to not meet eligibility requirements for the study after session 1 and their data was not analyzed. Twenty-four did not reach criterion performance in the generalization phase (see above) in session 1, and were invited to complete the second and third sessions in behavioral testing settings rather than in the scanner. Six of these participants dropped out of the study before completing both remaining sessions and two did not have complete data due to computer errors. Twenty-two participants passed the generalization criterion and were asked to complete the remaining two sessions in the scanner. Of these participants, two had very low accuracy in the generalization phase in the fMRI sessions, indicating a failure to learn the task structure, and their data was not analyzed. Two further participants were excluded because of excessive movement (more than 1 voxel) in more than one run. One participant had excessive movement only in the last run of the last session and this single run was discarded from analysis. One participant could not complete the experiment due to problems with the scanner on the day of testing, and one participant dropped out of the study before finishing both scanned sessions. In total, 16 participants (10 female, mean age 21.3 years, SD = 3.3 years) passed the generalization criterion on the first day, completed both days of the fMRI experiment and were included in analyses, while 16 participants (11 female, mean age 22.3 years, SD = 2.5 years) did not pass this criterion on the first day and completed the

remaining two days of testing in behavioral sessions. All participants gave informed consent as approved by the Human Research Protections Office at Brown University and were compensated for their participation.

Materials

Eighteen images of natural scenes from Konkle et al. (2010) [22] were used to represent contexts within kingdoms (nine for session 1 and nine for sessions 2 and 3). These scenes were chosen to be distinct from each other in content and did not include any visible animals, people or man-made objects (i.e. the image categories included in the fMRI task). Over the course of the experiment, participants saw 360 images in the initial category training for each image category (randomly sampled from a larger set of 468 images). Hand images were taken from the 11k Hands Database [23], leaf images from the Leafsnap database [24], and food images from the Bank of Standardized Stimuli (BOSS) [25], as well as the Foodpics database [26]. In the new category training and generalization phases, 546 images were used from each stimulus category. These included faces from the Chicago Face Database [27], animals from the BOSS and CARE databases [28], and man-made objects, also from the BOSS database. All category images were cropped or padded with white pixels to fit within a square image with a white background.

fMRI acquisition procedures

Whole-brain imaging was acquired using a Siemens 3T Magnetom Prisma system with a 64-channel head coil. In each fMRI session, a high resolution T1 weighted MPRAGE image was acquired for visualization (repetition time (TR), 1900 ms; echo time (TE), 3.02 ms; flip angle, 9°; 160 sagittal slices; 1 x 1 x 1 mm voxels). Functional volumes were acquired using a gradient-echo echo planar sequence (TR, 2000 ms; TE, 25 ms; flip angle 90°; 40 interleaved axial slices tilted approximately 30° from the AC-PC plane; 3 x 3 x 3 mm voxels). Functional data were acquired over three runs. Each run lasted 15.1 min on average (452 acquisitions). After the functional runs, a brief in-plane anatomical T1 image was collected, which was used to define a brain mask that respected the contours of the brain and the space of the functional runs (TR, 350 ms; TE 2.5 ms; flip angle 70°; 40 axial slices; 1.5 x 1.5 x 3 mm voxels). The sequence of scans was identical on both sessions. Soft padding was used to restrict head motion throughout the experiment. Stimuli were presented on a 32-inch monitor at the back of the bore of the magnet, and participants viewed the screen through a mirror attached to the head coil. Participants used a five-button fiber optic response pad to interact with the experiment (Current Designs, Philadelphia, PA).

fMRI preprocessing and analysis

Functional data were preprocessed using SPM12. Quality assurance for the functional data of each participant was first assessed through visual inspection and TSdiffAna (https://sourceforge.net/projects/spmtools/) and ArtRepair (http://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html). Outlier volumes (defined by standard deviations from the global average signal) were interpolated when possible. If interpolation was not possible, a nuisance regressor was added to the run with a stick function at the time points for these volumes. Slice timing correction was carried out by resampling slices to match the first slice. Next, motion during functional runs and days

was corrected by registering volumes to the first volume in the first session using rigid-body transformation.

A deformation matrix for spatial normalization to Montreal Neurological Institute (MNI) stereotaxic space using 4th order B-spline interpolation was calculated for the motion corrected functional volumes. The in-plane T1 anatomical image was used to create a brain mask for functional analysis using the Brain Extraction Tool in FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). This mask was then normalized to MNI space using the inverse deformation matrix from the normalization of the functional data.

Functional data were analyzed under the assumptions of the GLM using SPM12. Separate regressors were included for correct, erroneous and missed responses for each condition with the duration of each response set to participants trial-wise RT (or the duration of the stimulus display for missed responses). Nuisance regressors for participant motion (six translational and rotational components) were also included, as was an additional regressor for scan session. Regressors and parametric modulators were convolved with the SPM canonical hemodynamic response function (HRF). Functional data were pre-whitened and high-pass filtered at 0.008 Hz.

Representational similarity analysis

Whole-brain searchlight and ROI-based RSA were carried out in a two-step process.

First, in participant-level analyses, an empirical RDM was estimated from the cross-validated Mahalanobis distances of the regressors for the 18 conditions averaged over runs (6 contexts x 3 categories) within the generalization phase within a designated subset of voxels in each participant (either defined by an ROI or a spherical searchlight) using the RSAtoolbox v.2.0 [29,30]. The lower triangle of this empirical RDM was extracted giving the distances for the 153 condition pairs in this multivoxel space.

Hypothesis RDMs were constructed based on the similarities/dissimilarities that would be expected for a pure representation of a given type. These included five main effects (latent state, context, value, image category and subjective visual similarity of contexts), as well as five interactions (latent state X category, latent state X value, value X category, context X category, context X value).

In the case of latent state, value and context RDMs, the hypothesized distances were simply set so that conditions that were the same on these factors would have smaller expected distances, and larger distances where these conditions were not the same. For the category RDM, hypothesized distances were based on asymmetric distances related to the perceptual-semantic distances of animate and inanimate stimuli observed in many other studies [31,32]. Namely, that animate categories (faces and animals) were expected to be more similar to each other than inanimate objects, but inanimate objects were expected to be more similar to animals than to faces. Distances for these hypothesis RDMs were set as ordinal integer numbers to reflect predicted distances (e.g. for the category RDM, the expected distance between two conditions with faces would be 1, the distance between faces and animals would be 2 and between faces and objects would be 4). The RDM for the subjective visual similarity of contexts was specific to each participant and derived from the Euclidean distances of these natural scene images in the similarity judgment task completed at the end of the third session of the experiment. Interaction RDMs were created by extracting main effect RDMs below the diagonal (i.e. the lower triangle of the RDM matrix) as a vector, mean-

centering these values and multiplying them together. These interaction regressors allowed us to test where pattern-similarity reflected the modulation of one task component another (e.g. where items belonging to different categories were represented more similarly because of shared value associations).

Second, these hypothesis RDMs were related to the empirical RDM through multiple linear regression analysis, where a coefficient was estimated relating each hypothesis RDM to empirical RDMs allowing us to parcel out variance in representational distances due to multiple factors in the same model [33]. The lower-triangle of all hypothesis RDMs were extracted as vectors, mean centered, and included as predictors, along with an intercept term, in a multiple linear regression analysis to calculate beta coefficients relating each hypothesis RDM to the empirical RDM.

Searchlight analyses

Whole-brain analyses were carried out by passing a spherical searchlight with a radius of 9 mm over each voxel within participants' brain mask in native space. For each participant, beta coefficients for hypothesis RDMs were calculated at each step and averaged over searchlight passes for all voxels included in the searchlight to compute fixed-effects in a first-level analysis. This approach is similar to a common approach of assigning coefficients to a the central voxel of each searchlight and then smoothing these maps before second-level tests (e.g. [34]), but requires one less experimenter degree of freedom in defining the full-width half-maximum (FWHM) of the smoothing kernel. Group level analyses were conducted by normalizing participants' beta coefficient maps to MNI space using the deformation field from the normalization of participants' functional data and computing a one-sample t-test against zero. These volumes were kept in a 1.5 x 1.5 x 1.5 mm space and not resampled. Whole brain t-statistic maps were thresholded at a cluster defining threshold of $P < 0.001$ uncorrected, and then a cluster extent threshold (k) was set using non-parametric cluster-based permutation tests (10,000 permutations for each contrast) that controlled for multiple comparisons and tested statistical significance at $P < 0.05$ [35].

Region of interest analyses

We defined three *a priori* anatomical ROIs in this study based on prior work using the Automated Anatomical Labelling (AAL2) atlas [36]. First, the OFC ROI was given the same definition used by a study examining latent state representations in this region[37], which followed that of Kahnt et al. (2012) [38]. This definition included the bilateral combination of the following regions: the superior orbital gyri, middle orbital gyri, inferior orbital gyri medial orbital gyri, and rectal gyri. The VTC ROI was given the same definition used by a study that showed that pattern activity in this region differentiated between visual images based on animacy [31], excepting the bilateral parahippocampal gyri. This definition included the bilateral lingual gyri, fusiform gyri and inferior temporal cortices. The HPC ROI was defined as the bilateral hippocampi. These masks were warped into participants' native brain space using the inverse deformation matrix for all ROI-based analyses. We first conducted an RSA analysis using all voxels within these ROIs by calculating empirical RDMs for the cross-validated Mahalanobis distances from voxels within these ROIs in the same way as in the searchlight analyses, and then using the same multiple linear regression model to

relate hypothesis RDMs to these empirical RDMs. These beta-coefficients were then subjected to a second-level, one-sample t-test against zero to estimate statistical significance.

These ROIs were also used as an explicit mask in searchlight analyses to test for effects that were below threshold at a whole-brain cluster-corrected level within HPC and OFC. For these analyses, statistical significance was evaluated as in the whole-brain searchlight analysis with cluster-based permutation tests to compute a cluster extent threshold within these smaller volumes for each contrast, controlling for multiple comparisons at $P < 0.05$.

Univariate analyses

Functional volumes were normalized to a 1.5 x 1.5 x 1.5 mm MNI space and smoothed with an 8 mm FWHM Gaussian isotropic kernel. Beta coefficients for single subject effects were estimated using a fixed-effects model in a first-level analysis. For whole-brain contrasts, these estimates were then included in a second-level analysis using a one-sample t-test against zero at each voxel. As with whole-brain RSA analyses, cluster-based permutation tests used to compute a cluster extent threshold controlling for multiple comparisons at $P < 0.05$.

Comparison with functional networks

The results of the whole-brain latent state RSA searchlight was compared with a cortical parcellation based on resting state functional connectivity data from 1000 individuals by Yeo et al. [11]. To find the degree of overlap within each of these functional networks, we calculated the proportion of voxels in the cluster-corrected latent state statistical map that fell within these 17 networks projected to MNI152 space and defined with liberal boundaries around the cortex.

Comparison across sessions

To assess the stability of task-relevant representations across scanning sessions, we carried out two separate searchlight RSA analyses using data from within each session, as above. The resultant coefficient maps from the multiple linear regression analysis were then contrasted between sessions in both directions using paired t-tests. These contrasts were then assessed with cluster-based permutation tests controlling for multiple comparisons at $P < 0.05$.

Methods references

21. Greve, D. N. optseq2. *Optseq Home Page* (2002).
22. Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychol. Sci.* (2010). doi:10.1177/0956797610385359
23. Afifi, M. 11K Hands: Gender recognition and biometric identification using a large dataset of hand images. *Multimed. Tools Appl.* (2019). doi:10.1007/s11042-019-7424-8
24. Kumar, N. *et al.* Leafsnap: A computer vision system for automatic plant species identification. in *Lecture Notes in Computer Science (including subseries Lecture Notes*

in Artificial Intelligence and Lecture Notes in Bioinformatics) (2012). doi:10.1007/978-3-642-33709-3_36

25. Brodeur, M. B., Guérard, K. & Bouras, M. Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS One* (2014). doi:10.1371/journal.pone.0106953
26. Blechert, J., Meule, A., Busch, N. A. & Ohla, K. Food-pics: an image database for experimental research on eating and appetite. *Front Psychol* **5**, 617 (2014).
27. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* (2015). doi:10.3758/s13428-014-0532-5
28. Russo, N. *et al.* Validation of the C.A.R.E. stimulus set of 640 animal pictures: Name agreement and quality ratings. *PLoS One* (2018). doi:10.1371/journal.pone.0192906
29. Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Comput Biol* **10**, e1003553 (2014).
30. Walther, A. *et al.* Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* (2016). doi:10.1016/j.neuroimage.2015.12.012
31. Chikazoe, J., Lee, D. H., Kriegeskorte, N. & Anderson, A. K. Population coding of affect across stimuli, modalities and individuals. *Nat Neurosci* **17**, 1114–1122 (2014).
32. Thorat, S., Proklova, D. & Peelen, M. V. The nature of the animacy organization in human ventral temporal cortex. *Elife* (2019). doi:10.7554/eLife.47142
33. Nassar, M. R., McGuire, J. T., Ritz, H. & Kable, J. W. Dissociable forms of uncertainty-driven representational change across the human brain. *J. Neurosci.* (2019). doi:10.1523/JNEUROSCI.1713-18.2018
34. Devereux, B. J., Clarke, A., Marouchos, A. & Tyler, L. K. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J. Neurosci.* (2013). doi:10.1523/JNEUROSCI.3809-13.2013
35. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* **15**, 1–25 (2002).
36. Rolls, E. T., Joliot, M. & Tzourio-Mazoyer, N. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage* (2015). doi:10.1016/j.neuroimage.2015.07.075
37. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* **91**, 1402–1412 (2016).
38. Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J. & Haynes, J. D. Connectivity-based parcellation of the human orbitofrontal cortex. *J. Neurosci.* (2012). doi:10.1523/JNEUROSCI.0257-12.2012