

Current Biology, Volume 23

Supplemental Information

Decoding the Brain's Algorithm for Categorization

from Its Neural Implementation

Michael L. Mack, Alison R. Preston, and Bradley C. Love

Supplemental Information Inventory

Figure S1. Figure S1 extends Figure 2b to include data from a control analysis that investigated the role of complexity in our approach and a variant of the exemplar model that does not include selective attention.

Figure S2. Figure S2 extends Figure 2b by highlighting the brain regions that showed a significant difference between exemplar and prototype model-brain consistency when applying our method to a region of interest analysis.

Figure S3. Figure S3 extends Figure 3 to show the group-level maps of a searchlight analysis comparing the representational similarity of the prototype model to neural similarity.

Table S1. Table S1 lists the category structure of the stimuli used in the experiment by showing the value of the four dimensions for each stimulus.

Table S2. Table S2 lists all ROIs investigated in the ROI-based analysis briefly discussed in the main text and shown in Figure S2.

Supplemental Experimental Procedures

Supplemental References

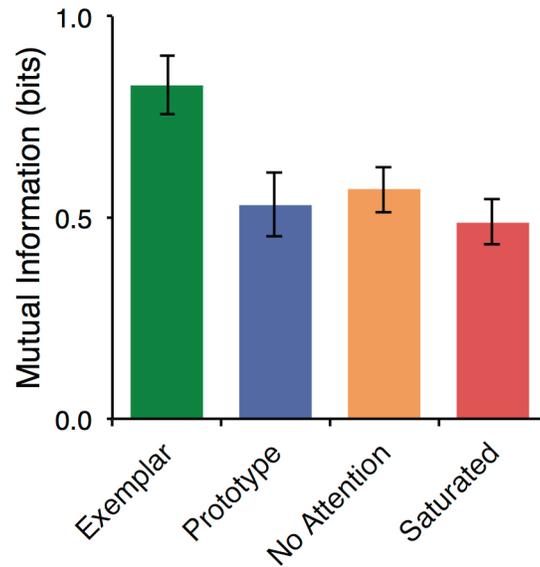
Figure S1, related to Figure 2

Figure S1. Model-brain consistency. The mutual information between responses of brain patterns as revealed by MVPA and representational match was higher for the exemplar relative to the prototype model, an over-parameterized saturated model, and a model with no attention weights. Data represent mean \pm 95% confidence intervals of the within-subject comparison between models.

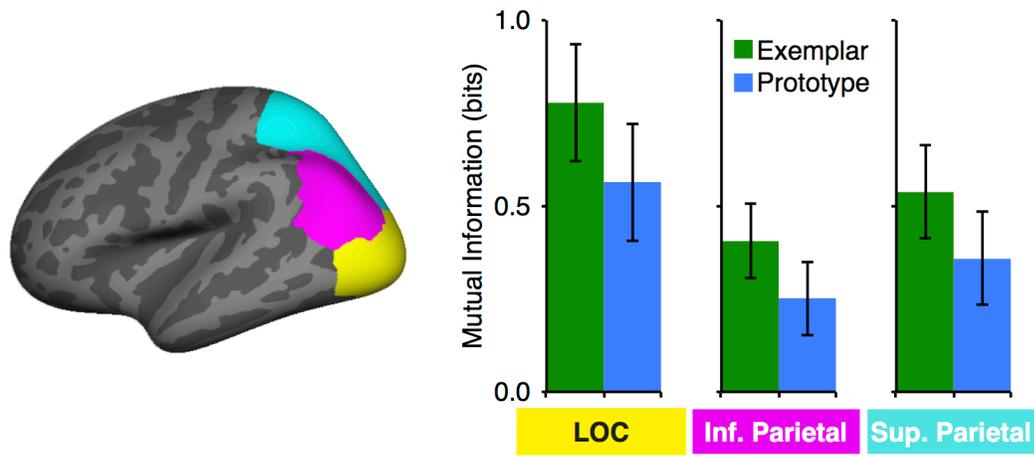
Figure S2, related to Figure 2

Figure S2. ROI analyses. Three anatomically-defined ROIs (highlighted in the brain on the left) showed significantly more consistency between brain response and exemplar model predictions than prototype model predictions: lateral occipital cortex (yellow), and inferior (magenta) and superior parietal lobules (cyan). Data represent mean \pm 95% confidence intervals ($\alpha=0.0015$).

Figure S3, related to Figure 3

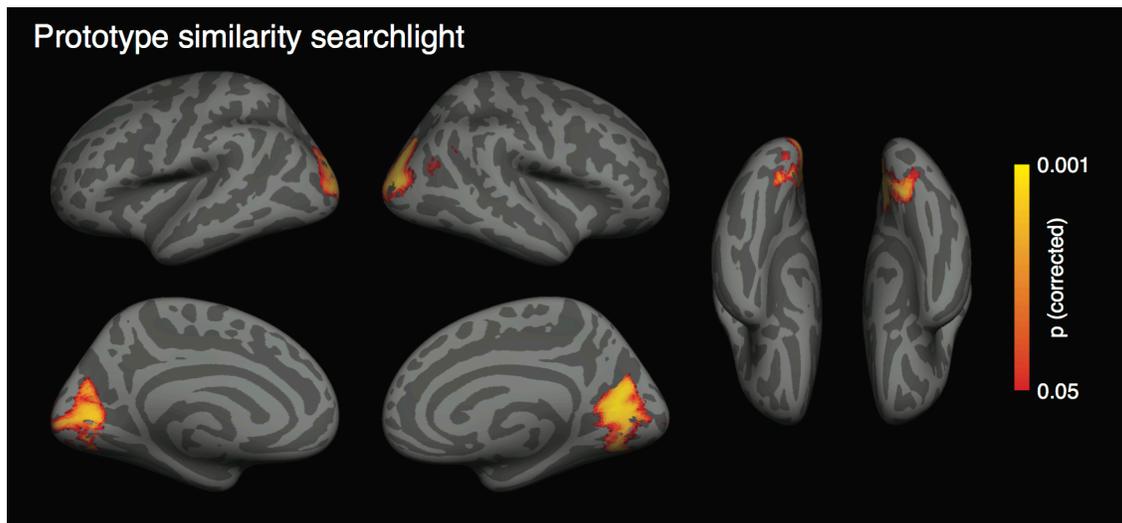


Figure S3. Representational similarity analysis of prototype representations. Searchlight analyses revealed correspondence between the similarity of neural representations and pairwise similarity predictions from the prototype model for the objects was restricted to early visual areas ($p < 0.05$, FWE corrected).

Table S1, Related to Figure 1

Stimulus	Dimension			
	1	2	3	4
Category A				
A1	1	0	0	0
A2	1	0	1	0
A3	0	1	0	0
A4	0	0	1	0
A5	0	0	0	1
Category B				
B1	1	1	0	0
B2	1	0	0	1
B3	0	1	1	1
B4	1	1	1	1
Transfer				
T1	0	1	1	0
T2	1	1	1	0
T3	0	0	0	0
T4	1	1	0	1
T5	0	1	0	1
T6	0	0	1	1
T7	1	0	1	1

Table S1. 5/4 category structure.

Table S2, related to Figures 2 and S2

Region of Interest	Model brain consistency (MI)	
	Exemplar	Prototype
Lateral occipital cortex*	13 (0.779)	4 (0.565)
Caudate	5 (0.268)	5 (0.238)
Inferior frontal gyrus	5 (0.276)	4 (0.241)
Hippocampus	7 (0.310)	1 (0.254)
Medial temporal cortex	6 (0.273)	3 (0.244)
Pericalcarine	11 (0.640)	4 (0.455)
Lingual gyrus	5 (0.483)	5 (0.455)
Cuneus	8 (0.531)	3 (0.449)
Fusiform	6 (0.365)	5 (0.317)
Inferior parietal lobule*	11 (0.406)	3 (0.253)
Inferior temporal cortex	5 (0.278)	4 (0.265)
Lateral orbital frontal cortex	3 (0.259)	2 (0.257)
Medial orbital frontal cortex	4 (0.247)	6 (0.260)
Frontal pole	6 (0.264)	6 (0.254)
Rostral medial frontal cortex	5 (0.362)	2 (0.300)
Superior prefrontal cortex	7 (0.352)	3 (0.284)
Retrosplenial cortex	9 (0.296)	4 (0.230)
Caudal anterior cingulate cortex	4 (0.260)	7 (0.280)
Putamen	7 (0.267)	1 (0.241)
Amygdala	9 (0.262)	6 (0.236)
Medial temporal cortex	6 (0.267)	4 (0.243)
Paracentral gyrus	5 (0.291)	1 (0.223)
Posterior cingulate cortex	5 (0.290)	4 (0.259)
Precentral gyrus	9 (0.299)	0 (0.220)
Postcentral gyrus	8 (0.356)	2 (0.270)
Precuneus	9 (0.438)	4 (0.325)
Rostral anterior cingulate cortex	7 (0.285)	3 (0.242)
Superior parietal lobule*	11 (0.537)	2 (0.358)
Superior temporal gyrus	2 (0.226)	7 (0.259)
Supramarginal gyrus	5 (0.298)	4 (0.256)
Temporal pole	3 (0.249)	5 (0.272)
Transverse temporal gyrus	5 (0.284)	4 (0.255)
Insular cortex	5 (0.253)	3 (0.240)

Table S2. Results of ROI analysis. The number of participants with significantly greater than chance brain-model consistency is reported for each model. Mean MI across subjects is reported in parentheses. Regions marked with an asterisk (*) showed significantly higher MI for exemplar model consistency than prototype model consistency. No regions showed significantly higher MI for the prototype model.

Supplemental Experimental Procedures

Participants. Twenty-three participants participated in the experiment. Two participants were removed prior to analysis for excessive head motion during fMRI scanning and one participant was removed for failure to learn the categorization task. The remaining twenty participants were included in the primary analysis (ages range of 19-33 years; mean age of 23.5 years; 14 female).

Stimuli. The stimulus set [S1] was constructed by including sixteen objects consisting of simple shape enclosed in a grey horizontally oriented rectangle (Figure 1). The simple shape varied based on four salient binary-valued features (color: red or green, shape: circle or triangle, size: large or small, and position: right or left). For each participant, the four features were randomly assigned to the four dimensions defined by the 5/4 category structure. This structure is divided into two categories with the logical values of the prototype member of category A corresponding to [0,0,0,0] and the logical values of the prototype member of category B corresponding to [1,1,1,1]. Nine objects served as the training items with five for category A and four for category B. The remaining seven objects served as a transfer set (Table S1).

Procedures. After an initial screening and consent in accordance with the University of Texas Institutional Review Board, participants were instructed on the category learning task. These instructions explained that the participant would be shown simple objects composed of different features and that the task was to learn which object belonged to one of two categories through corrective feedback.

Participants performed the training phase of the experiment in a behavioral testing room on laptop computer. On each training trial, one of the nine training stimuli was displayed for 3.5s and participants made a response to the stimulus's category by pressing one of two labeled keys on the keyboard. Then, a fixation cross was presented for 0.5s, followed by a feedback display that presented the stimulus, the correct category, and whether the participant's response was correct or incorrect for 3.5s. Trials ended with a 0.5s fixation cross. The nine training stimuli were presented 20 times in randomized order during the initial training outside the scanner. Participants also completed additional training trials inside the fMRI scanner during an anatomical scan as a refresher of the training items' category membership. In total, across the entire training phase, participants completed 24 repetitions with each training stimulus.

After training, participants performed the testing phase during functional scanning. On each test trial, one of sixteen stimuli (consisting of the nine training stimuli and seven novel transfer stimuli) was displayed for 3.5s and participants made a category response by pressing one of two buttons on a MRI-compatible button box. A fixation cross was then presented for 6.5s. No feedback was provided during the testing phase. The sixteen stimuli were presented three times in randomized order during six functional runs for eighteen total repetitions per stimulus.

Modeling Procedures. The exemplar model [S2] posits that categories are represented by storing individual exemplars in memory. The exemplars are represented as points in a multidimensional space. The psychological distance between two stimuli is modeled as a weighted city-block metric,

$$d_{ij} = \sum_{m=1}^4 w_m |x_{im} - x_{jm}|,$$

where x_{im} is the value (0 or 1) of item i on dimension m , and w_m ($0 \leq w_m \leq 1$, $\sum w_m = 1$) is the attention weight for dimension m . The similarity between two stimuli is an exponential decay function of their distance,

$$s_{ij} = e^{-cd_{ij}},$$

where c is an overall sensitivity parameter. The probability that item i is categorized into category A is given by,

$$P(A|i) = \frac{\left[\sum_{a \in A} s_{ia} \right]^\gamma}{\left[\sum_{a \in A} s_{ia} \right]^\gamma + \left[\sum_{b \in B} s_{ib} \right]^\gamma},$$

where γ is a response scaling parameter such that when $\gamma = 1$, responses match the relative summed similarity and when $\gamma > 1$, responses are more deterministic. The full version of the exemplar model has five free parameters: sensitivity c , response scaling γ , and three freely varying attention weights.

The prototype model [S3, S4] posits that categories are represented by prototype representations. The distance between item i and Prototype A is given by the city-block metric,

$$d_{iA} = \sum_{m=1}^4 w_m |x_{im} - x_{Am}|,$$

where x_{im} is the value (0 or 1) of item i on dimension m , x_{Am} is the value of Prototype A on dimension m , and w_m are the attention weights as in the exemplar model. The similarity between item i and Prototype A is given by,

$$s_{iA} = e^{-cd_{iA}},$$

where c is an overall sensitivity parameter. The probability that item i is categorized into category A is given by,

$$P(A|i) = \frac{s_{iA}^\gamma}{s_{iA}^\gamma + s_{iB}^\gamma},$$

where γ is a response scaling parameter. However, in this form of the prototype model, γ cannot be estimated separately from the sensitivity parameter c , thus γ was set to 1. The prototype model has four free parameters: sensitivity c and three freely varying attention weights.

For each participant, the exemplar and prototype models were fit to the response probabilities from the last half of the training phase. Model parameters were optimized with standard maximum likelihood techniques [S5]. Exemplar and prototype model fits

were compared for each participant with a chi-square test of the G likelihood ratio statistic.

Model measures of representational match were obtained for both models using the models' optimized parameter sets. Representational match is calculated as the summed similarity of a test object to all the stored category representations from a model. Thus, for the exemplar model, the representational match for item i is given by,

$$rm_i = \sum_{a \in A} s_{ia} + \sum_{b \in B} s_{ib} ,$$

and for the prototype model,

$$rm_i = s_{iA} + s_{iB} .$$

For display purposes (see Figure 2a), representational match values were normalized such that the maximum value for a model was equal to 1 and minimum value equal to 0. Following standard machine learning practices, representational match values were z-scored prior to MVPA.

fMRI data acquisition. Whole-brain imaging data were acquired on a 3.0T GE Signa MRI system (GE Medical Systems). Structural images were acquired using a T2-weighted flow-compensated spin-echo pulse sequence (TR=3s; TE=68ms, 256x256 matrix, 1x1mm inplane resolution) with thirty-three 3-mm thick oblique axial slices (0.6mm gap), approximately 20° off the AC-PC line. Functional images were acquired with an echo planar imaging sequence using the same slice prescription as the structural images (TR=2s, TE=30.5ms, flip angle=73°, 64x64 matrix, 3.75x3.75 in-plane resolution, bottom-up interleaved acquisition, 0.6mm gap). An additional high-resolution T1-weighted 3D SPGR structural volume (256x256x172 matrix, 1x1x1.3mm voxels) was acquired for registration and cortex parcellation.

fMRI data preprocessing. Data were preprocessed and analyzed using FSL4.1 [S6] and custom Matlab routines. Functional images were realigned to the first volume in the time series to correct for motion, coregistered to the T2-weighted structural volume, high-pass filtered (128s), and detrended to remove linear trends within each run. All analyses were performed in the native space of each participant.

Multivoxel pattern analysis. Pattern classification analyses were implemented using PyMVPA [S7] and custom Python code. The goal of the MVPA was to assess the extent that coherent activation patterns could predict trial-by-trial values of representational match from the exemplar and prototype models. MVPA performance for each model was assessed in separate analyses. Each trial was labeled with the corresponding representational match value for the stimulus presented. To account for the hemodynamic lag, volumes from 4, 6, and 8 seconds after stimulus onset were averaged for each trial. A MVPA regression method based on linear support vector regression (SVR) was trained and evaluated with leave-one-out six-fold cross-validation using the six functional runs. Cross validation was used to ensure the SVR solution generalized across independently sampled functional runs and was not the result of overfitting. On each fold, four runs served as the training data and two runs were used

to evaluate the training. Within each cross validation fold before SVR training, feature selection was performed on the training dataset. Feature selection consisted of conducting a univariate contrast of task greater than baseline across the whole brain and retaining the 1000 voxels with the highest F values for MVPA training and testing. One of the test runs was used to optimize the SVR parameters C and ϵ , both of which determine the complexity of the SVR estimation. The second test run was used to evaluate the resulting trained SVR with optimal parameters. SVR output of representational match for each of the sixteen stimuli cross validation folds served as the prediction of representational match from coherent patterns of brain activation.

Evaluating model consistency with brain response. Existing model-based methods of fMRI analysis [S8] offer important advances that allow one to map out voxel activations that track model measures. However, these techniques are ultimately limited for comparing model consistency to brain response as they depend on correlations between model measures and activation in individual voxels (considered independently of one another). For any given voxel, one model may correlate more with activation than another model, but to date, the extension of these results to regions or the whole brain has been elusive. In practice, model selection at the level of individual voxels depends on the relative number of voxels tracking model predictions [S9, S10]. The more voxels tracking a model, the more it is favored. But, when testing the predictions of two models, the goal is not to characterize the response of each voxel in the brain; rather the goal is to determine overall what model provides a more accurate description of the brain responses (either globally or within specific regions). Determining fit by the spatial extent (number of voxels) could be misleading. For example, a model with many weakly correlated voxel activations may be selected, when, in fact, the predictions from a competing model are perfectly tracked by activations in relatively fewer voxels.

Information pattern analyses, such as MVPA [S11] and RSA [S12], take the crucial step of reversing the inference of standard GLM analyses. In this type of analysis, the question is no longer what conditions predict voxel activation, but what patterns of activation contain information consistent with experimental conditions. However, to date, the application of pattern analyses has been limited to decoding representations of different kinds of stimuli [S13, S14] or different types of experimental conditions [S11, S15].

We assessed the consistency between brain and model by marrying these two breakthroughs (model-based fMRI analyses and MVPA) and grounding them in a model selection framework. The consistency between brain and model was assessed by a comparison of a model's representational match and the SVR prediction of those values. This consistency was evaluated with mutual information. Mutual information (MI) measures the mutual dependence between two random variables X and Y :

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where $p(x,y)$ is the joint probability distribution function of X and Y and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y . In the current case, the random variables of interest were the model and brain predictions of representational match. Marginal and joint probability distribution functions for model and brain predictions of representational match were estimated by a smooth bootstrapping procedure where a small amount of zero-centered random noise ($\sigma = 1/500$) was added to 500 samples from the representational match predictions, a procedure akin to kernel density estimation. For the brain, these predictions consisted of the SVR results of representational match from the cross validation folds. For the models, these predictions consisted of representational match predictions derived from model parameters from multiple iterations of model fits to randomly selected subsets of training data. The resulting probability distribution functions were discretized and mutual information was calculated with the equation presented above [S16].

Our choice of mutual information was two fold. First, mutual information measures the amount of information shared between two functions while taking into account the variability of those functions. This is critical in that it is likely that latent measures across different models will vary in their entropy. For example, as can be appreciated in Figure 2, representational match for the exemplar model is more variable than for the prototype model. This is an issue since the more variable a model measure, the more difficult it is to predict that measure with MVPA. Thus, it is key that a measure of consistency between model and brain predictions takes this into account by scaling by the variability of the two functions. Second, unlike correlation, MI is not limited to linear relationships or assumptions of normality. Although these assumptions may hold in some cases, making correlation a valid measure of consistency, nonlinearities with brain response and model measures are likely, thus warranting use of mutual information.

A bootstrapping procedure was used to evaluate greater than chance consistency between brain and model and for model comparison. This procedure consisted of randomly shuffling the stimulus labels of the training data, performing the MVPA analysis detailed above, and calculating mutual information. This procedure was repeated 1000 times to create a null distribution for each participant. Mutual information scores greater than 95% of the null distribution values were considered significantly greater than chance. Also, for each participant, the difference between the randomly shuffled exemplar and prototype mutual information scores served as a null distribution for comparing model consistency. A difference between a participant's exemplar and prototype model mutual information greater than 95% of this null distribution were considered significantly different.

Model recovery simulations. To validate our approach, we performed the MVPA plus MI analysis on activation patterns from 5000 simulated voxels. Activation profiles for these simulated voxels were constructed as follows. Average parameter values for both exemplar and prototype models were calculated from the behavioral fits explained above. In separate simulations, a percentage (1%, 5%, 10%, 25%) of the simulated voxels were assigned activation profiles equal to the stored model exemplars

or prototypes following the parameters of the two models. The remaining voxels had activation profiles randomly selected from a Gaussian noise distribution with a mean and standard deviation equal to the mean and standard deviation of the dimensional values from the exemplars and prototypes. Cross-validated MVPA was performed on the simulated datasets to extract the embedded representational match from the two models. The consistency between the MVPA output and the models' representational match values was assessed with mutual information as described above. Over 1000 repetitions of each simulation, brain response was always more consistent with the generating model than the other model.

Saturated model. We used mutual information as an index of model-brain consistency. To validate this choice, we evaluated the consistency between brain response and a saturated version of the exemplar model that included separate memory strength parameters for each of the stored training items for a total of 14 free parameters. The logic of this control analysis was to purposefully use a model that overfits the behavioral data to derive representational match values and attempt to predict this overfitted model's latent measure with brain states. A model that overfits behavioral data is not capturing an underlying psychological process, but is instead fitting noise. Because of this, an overfitted model's internal state is unlikely to capture changes in brain state. We predicted that the overly complex saturated model would provide an excellent fit to the behavioral data, but have low MI as its internal state would reflect noise in the behavioral data as opposed to a change of state in the brain. Such a result would demonstrate that absolute behavioral fit and MI need not correspond. Across the participants, the saturated model provided a better fit numerically (2.45 mean fit error) to the categorization response probabilities than both the exemplar (4.94 mean fit error) and prototype (6.04 mean fit error) models (it should be noted that if model complexity is taken into account, the saturated model fits the behavioral data significantly worse than the exemplar and prototype models). However, the MVPA plus MI results showed significantly worse consistency between the saturated model and the exemplar model (Figure S1). With too many free parameters, the saturated model overfit the behavioral responses resulting in unstable and overly complex predictions of representational match. This unjustified complexity is penalized by the mutual information index of model-brain consistency. Thus, by including brain response in model selection the more consistent and parsimonious exemplar model is selected.

Representational similarity analysis. Representational similarity analysis (RSA) was performed to compare the similarity structure of the stimuli as predicted by the exemplar model to the similarity of the neural patterns for the stimuli [S12]. First, the average voxelwise activation for each of the stimuli in each run was estimated with a GLM that included separate regressors for the stimuli plus confound regressors for the motion parameters. The resulting beta images for the stimuli were averaged over the six runs to give the average voxelwise activation across the brain for each stimulus. These average beta images were submitted to a searchlight analysis (sphere radius of 5 mm) to compare the similarity between the neural patterns of the stimuli to the similarity as

predicted by different models. Exemplar model dissimilarity matrices (DM) were derived for each participant by taking the Pearson correlation distance between all pairs of the stimuli with the feature values of the stimuli weighted according to the attention weights from the model fits to training behavior. Neural DMs were calculated within each searchlight sphere by taking the Pearson correlation distance between all pairs of stimulus beta patterns. For each searchlight, the correspondence between the exemplar model DM and neural DM was evaluated with Spearman correlation and recorded at the searchlight center. The same procedure was used to compare the physical similarity of the stimuli, i.e. a model with no attention weighting, to neural pattern similarity. In this analysis, physical similarity DMs were calculated by taking the non-weighted Pearson correlation distance between all pairs of the stimuli.

The statistical significance of the RSA correlation values was evaluated with a Monte Carlo method by randomly permuting the labels of the neural DM and calculating the correlation between the model and neural DMs. This was repeated 1000 times to create a distribution of chance correlations. The observed correlation was compared to the chance correlation distribution to create p maps for each participant. These p maps were converted to z -scores. For group analysis, voxelwise nonparametric permutation testing (5000 permutations) was performed using Randomise in FSL [S6]. The resulting statistical maps were family-wise error corrected using $P < 0.05$, based on the threshold-free cluster enhancement statistic image [S17].

Region of interest analysis. A cortical parcellation of the high-resolution T1 SPGR volume was obtained for each participant using FreeSurfer (Martinos Center for Biomedical Imaging, MGH, Charlestown, MA). From these parcellations, masks for thirty-three bilateral ROIs (Table S2) were created for ROI-based MVPA analyses. The same MVPA plus MI approach described in the whole brain analysis was used for the ROI analyses with the exception that only the voxels within an ROI mask were used for MVPA. Consistency between brain and model within ROIs was assessed in two ways. First, the mean MI scores for both models within each ROI were compared across participants with paired t -tests (FDR corrected for multiple comparisons, Figure S2). Second, the bootstrapping procedure explained in the whole brain analysis was applied to all ROI results to assess the consistency between both models and the brain response within the ROI. The number of participants showing greater than chance consistency between brain and model for each ROI are reported in Table S2.

Supplemental References

- S1. Medin, D. L., and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review* 85, 207–238.
- S2. Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115, 39–61.
- S3. Minda, J. P., and Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 28, 275–292.
- S4. Zaki, S. R., Nosofsky, R. M., Stanton, R. D., and Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory and Cognition* 29, 1160–1173.
- S5. Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* 47, 90–100.
- S6. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage* 62, 782–90.
- S7. Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V, and Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7, 37–53.
- S8. O’Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences* 1104, 35–53.
- S9. Hampton, A. N., Bossaerts, P., and O’Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America* 105, 6741–6746.
- S10. Davis, T., Love, B. C., and Preston, A. R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex* 22, 260–273.
- S11. Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10, 424–430.

- S12. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2, 4.
- S13. Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8, 679–685.
- S14. Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929.
- S15. Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., and Haxby, J. V (2012). The representation of biological classes in the human brain. *Journal of Neuroscience* 32, 2608–2618.
- S16. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters* 80, 197–200.
- S17. Smith, S. M., and Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44, 83–98.