Supplementary Material for

Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members

Tyler Davis*, Bradley C. Love, Alison R. Preston

*To whom correspondence should be addressed. Email: thdavis@mail.utexas.edu

This file includes:

Tables S1 to S5
Supplementary Methods

Supplemental Tables

Table S1. Regions correlated with the recognition strength measure from SUSTAIN during the stimulus presentation period.

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Anterior Cingulate Cortex	R	3.24	12, 28, 26
Caudate	R	4.42	10, 10, -2
Cerebellum	L	4.05	-8, -22, -30
		4.40	-22, -58, -36
		3.59	0, -62, -34
		3.50	-6, -76, -24
Fusiform Gyrus	L	3.44	-22, -86, -6
	R	3.88	40, -58, -10
		3.25	38, -44, -18
Frontal Pole	L	3.52	-44, 40, 32
	R	4.26	50, 28, 24
Hippocampus	L	3.42	-24, -32, -6
	R	3.88	20, -38, 2
Inferior Frontal Gyrus—Pars Triangularis	L	3.90	-52, 24, 18
	R	3.62	48, 28, 6
Inferior Temporal Gyrus	L	4.25	-40, -58, -6
Insula	L	4.53	-32, 16, -10
Intracalcerine Cortex	L	3.71	-16, -72, 6
	R	3.33	14, -68, 16
Midbrain	Bilateral	5.40	0, -24, -24
Middle Frontal Gyrus	L	4.15	-36, 4, 42
		3.31	-38, 20, 28
	R	4.52	48, 14, 32
		4.03	46, 6, 52
		3.33	30, 2, 42
Lateral Occipital Cortex—Inferior Division	R	3.88	44, -86, 12
Lateral Occipital Cortex—Superior Division	L	4.50	-26, -60, 40
		3.81	-26, -76, 52
	R	4.28	30, -66, 32
		3.78	18, -62, 52
Nucleus Accumbens	L	4.44	-8, 12, -4
Occipital Pole	R	3.75	16, -88, 2
Orbital Frontal Cortex	R	4.87	34, 28, -6
Pallidum	R	4.19	12, -6, -6
Paracingulate Gyrus	L	3.73	-4, 10, 52
	R	4.27	4, 26, 44
Precuneus	R	4.56	14, -70, 36
Posterior Cingulate Gyrus	R	4.01	8, -18, 30
Supramarginal Gyrus	L	4.17	-42, -44, 44
Superior Frontal Gyrus	R	4.06	12, 12, 58
Superior Parietal Lobule	R	4.71	32, -50, 42
Thalamus	L	3.95	-12, -10, -4
	R	3.37	20, -24, 8

Table S2. Regions correlated with the error correction measure from SUSTAIN during the feedback period.

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Angular Gyrus	L	4.55	-62, -56, 14
Anterior Cingulate Cortex	R	4.05	2, 6, 26
Caudate	R	4.89	10, 6, 2
	L	3.32	-14, 8, 6
Cerebellum	R	3.95	6, -56, -38
	L	3.94	-18, -76, -28
	L	3.73	-30, -76, -28
Fusiform Gyrus	R	4.36	34, -48, -16
	L	4.12	-32, -50, -12
Frontal Pole	R	3.62	32, 48, 4
	L	3.42	-32, 54, 0
Hippocampus	R	4.76	24, -40, 0
Inferior Frontal Gyrus—Pars Opercularis	L	4.24	-44, 22, 20
Inferior Frontal Gyrus—Pars Triangularis	R	4.47	48, 26, 10
Insula	R	3.29	30, 12, -18
	L	4.83	-32, 18, -8
Juxtapositional Lobule	L	4.17	-14, 10, 46
Midbrain	R	5.00	4, -28, -6
Middle Frontal Gyrus	R	5.24	36, 20, 32
	L	4.08	-40, 12, 58
Middle Temporal Gyrus	R	3.36	72, -30, -4
	L	4.15	-50, -28, -6
Lateral Occipital Cortex—Inferior Division	L	3.36	-50, -69, 10
Lateral Occipital Cortex—Superior Division	R	5.4	24, -60, 50
		4.69	36, -74, 38
		3.76	26, 78, 50
	L	4.45	-34, -66, 60
		3.95	-22, -78, 54
		3.40	-28, -86, 54
Orbital Frontal Cortex	R	4.76	38, 22, -6
Paracingulate Gyrus	R	4.89	10, 38, 22
		3.89	14, 22, 38
Precuneus	R	4.63	14, -64, 34
Precentral Gyrus	L	4.55	-36, 6, 26
Posterior Cingulate Gyrus	R	3.99	8, -26, 26
	L	4.47	-8, -30, 26
Supramarginal Gyrus	R	5.07	42, -46, 40
Superior Frontal Gyrus	R	4.24	12, 16, 54
Superior Temporal Gyrus	L	3.70	-50, 2, -16
Superior Parietal Lobule	L	4.01	-32, -54, 38
Thalamus	R	3.99	10, -14, 8
	L	3.94	-10, -8, 0
		3.60	-10, -26, 6

Table S3. Regions correlated with the error correction measure of ALCOVE during the feedback period.

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Angular Gyrus	L	3.69	-40, -54, 24
Anterior Cingulate Cortex	R	4.39	4, 6, 26
		3.39	10, 44, 6
Caudate	R	4.20	10, 8, 2
Cerebellum	R	4.48	0, -58, -28
		3.49	30, -66, -28
		3.36	10, -76, -22
	L	4.70	-2, -46, -14
		4.03	-16, -36, -26
		4.43	-14, -74, -28
		4.03	-32, -74, -30
Fusiform Gyrus	R	4.96	34, -50, -18
		3.60	40, -12, -36
	L	4.67	-30, -52, -12
Frontal Pole	R	3.63	32, 48, 4
	L	3.71	-24, 52, 14
Hippocampus	R	4.46	24, -20, 0
Inferior Frontal Gyrus—Pars Opercularis	R	4.93	46, 24, 10
	L	4.70	-44, 22, 20
		4.19	-46, 24, -4
Inferior Occipital Gyrus	R	3.74	44, -80, 12
Inferior Temporal Gyrus	R	3.79	48, -54, -10
		3.45	56, -20, -32
Juxtapositional Lobule	L	4.68	-14, 8, 48
Midbrain	R	5.36	4, -28, -4
	L	4.57	-12, -16, -10
Middle Frontal Gyrus	R	5.57	34, 18, 34
	L	4.20	-42, 12, 58
Middle Temporal Gyrus	R	4.67	48, -18, -12
		3.91	62, -30, -6
	L	4.72	-50, -26, -6
Lateral Occipital Cortex—Inferior Division	L	3.92	-46, -78, 12
Lateral Occipital Cortex—Superior Division	R	5.32	22, -58, 50
		4.74	30, -76, 40
	L	4.43	-20, -72, 40
		3.99	-32, -62, 56
		3.47	-28, -86, 34
		3.23	-32, -84, 18
Orbital Frontal Cortex	R	4.91	40, 20, -8
	L	5.11	-24, 22, -10
Pallidum	L	4.68	-12, 2, 2
Paracingulate Gyrus	R	4.85	10, 40, 22
		4.55	8, 26, 34
	L	3.45	-8, 44, 12
Planum Polare	R	4.81	48, 0, -16
Precuneus	R	4.68	14, -66, 34
	L	3.92	-18, -60, -36

Table S3. Continued

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Precentral Gyrus	R	3.49	36, -8, 44
	L	4.85	-48, -4, -20
		4.67	-42, 0, 32
Posterior Cingulate Gyrus	R	4.25	2, -24, -22
		4.66	10, -26, 28
	L	4.90	-8, -28, 26
Supramarginal Gyrus	R	4.51	42, -44, 40
	L	3.37	-44, -34, 38
Superior Frontal Gyrus	R	4.75	12, 16, 56
	L	3.17	-24, -4, 54
		4.01	-4, 46, 32
Superior Temporal Gyrus	L	3.80	-66, -40, 4
Superior Parietal Lobule	L	3.69	-32, -54, 38
Temporal Pole	R	3.55	38, 4, -40
Thalamus	R	4.76	8, -14, 6

Table S4. Regions demonstrating greater activation for correctly categorized exception items relative to correct rule-following items during stimulus presentation.

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Anterior Cingulate Cortex	L	3.57	-8, 40, 16
Caudate	R	5.98	8, 8, -2
	L	5.93	-8, 12, -2
Cerebellum	R	4.74	36, -56, -32
		4.08	14, -50, -16
		3.97	26, -66, -22
	L	4.81	-26, -64, -30
		4.21	-6, -76, -24
		3.87	-4, -60, -34
		3.76	-42, -60, -38
		3.73	0, -66, -8
		3.53	-24, -46, -30
Fusiform Gyrus	R	4.18	40, -50, -10
	L	4.37	-20, -84, -8
		4.37	-30, -70, -10
Frontal Operculum Cortex	L	4.25	-42, 20, 6
Frontal Pole	R	4.72	48, 38, 24
		4.31	42, 58, 12
		3.92	26, 50, -16
		3.14	40, 52, -6
	L	3.73	-18, 50, -20
		3.40	-14, 66, -14
Hippocampus	R	4.67	26, -26, -8
		4.64	18, -36, 4
Inferior Frontal Gyrus—Pars Triangularis	R	4.09	48, 28, 8
	R	4.91	-58, 26, 20
Inferior Temporal Gyrus	L	4.48	-40, -58, -6
Insula	R	5.78	34, 24, -4
Intracalcerine Cortex	R	4.12	14, -78, 14
	L	4.65	-10, -78, 12
Midbrain	R	5.70	2, -20, -12
Middle Frontal Gyrus	R	5.18	44, 16, 30
		4.81	36, 6, 40
	L	5.29	-42, 4, 50
		5.10	-36, 8, 32
		4.65	-40, 30, 20
		4.37	-34, 2, 64
Middle Temporal Gyrus	R	3.34	62, -34, -10
	L	3.65	-54, -34, -14
Lateral Occipital Cortex—Inferior Division	R	3.92	36, -84, 4
Lateral Occipital Cortex—Superior Division	R	5.10	32, -66, 32
	L	5.72	-28, -62, 40
		3.77	-18, -64, 54
Lingual Gyrus	R	4.82	30, -56, 2
Occipital Pole	R	4.67	18, -90, 2
		3.71	6, -94, 14
	L	3.70	-12, -96, 0
Orbital Frontal Cortex	L	5.59	-32, 24, -6
Pallidum	R	4.75	22, -8, -4

Table S4. **Continued**

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Paracingulate Gyrus	R	5.88	6, 30, 40
		4.60	6, 10, 54
		4.41	10, 40, 20
Precuneus	R	5.29	16, -70, 44
	L	5.86	-6, -74, 42
Precentral Gyrus	R	3.57	32, -18, 52
	L	3.66	-58, 6, 38
Posterior Cingulate Gyrus	R	4.23	4, -30, 32
Superior Parietal Lobule	R	5.54	34, -50, 42
	L	5.38	-36, -46, 40
Temporal Pole	L	3.98	-54, 18, -12
Thalamus	L	4.05	-10, -20, 12
		3.24	-20, -32, -2

Table S5. Regions correlated with the recognition strength measure from SUSTAIN, controlling for the effect of reaction time.

Region	Hemisphere	Z	MNI coordinates (x, y, z)
Caudate	R	4.42	10 , 10, -2
Cerebellum	L	4.05	-8, -22, -30
	L	4.40	-22, -58, -36
	L	3.61	-32, -62, -32
	L	3.50	-6, -76, -24
Frontal Pole	L	3.52	-44, 40, 32
	R	4.26	50, 38, 24
Fusiform Gyrus	R	3.88	40, -58, -10
	R	3.44	32, -52, -16
	R	3.25	38, -44, -18
	L	3.44	-22, -86, -6
Hippocampus	R	3.88	20, -38, 2
	R	3.77	26, -28, -8
	L	3.42	-24, -32, -6
Inferior Frontal Gyrus—Pars Triangularis	R	3.62	48, 28, 6
	L	3.90	-52 , 24, 18
Inferior Temporal Gyrus	L	4.25	-40, -58, -6
	L	3.95	-22, 46, -28
Insula	L	4.53	-32, 16, -10
Intracalcerine Cortex	R	3.24	18, -78, 6
	L	3.71	-16, -72, 6
	L	3.48	-10, -78, 14
	R	3.33	14, -68, 16
Lateral Occipital Cortex—Inferior Division	R	3.88	44, -86, 12
Lateral Occipital Cortex—Superior Division	L	4.50	-26, -60, 40
	R	4.28	30, -66, 32
	L	3.81	-26, -76, 52
	R	3.78	18, -62, 52
Lingual Gyrus	R	3.51	30, -58, 0
Midbrain	Bilateral	5.40	0, -24, -14
Middle Frontal Gyrus	L	4.15	-36, 4, 42
	L	3.31	-38, 20, 28
	R	4.52	48, 14, 32
	R	4.12	36, 6, 36
	R	4.03	46, 6, 52
	R	3.33	30, 2, 42
Nucleus Accumbens	L	4.44	-8, 12, -4
Occipital Pole	R	3.75	16, -88, 2
Orbital Frontal Cortex	R	4.87	34, 28, -6
Pallidum	R	4.19	12, -6, -6
Paracingulate Gyrus	R	4.27	4, 26, 44
	L	3.73	-4, 10, 52
	L	3.58	-8, 18, 48
Posterior Cingulate Gyrus	R	3.99	8, -18, 30
Precuneus	R	4.56	14, -70, 36
Superior Frontal Gyrus	R	4.06	12, 12, 58
Superior Parietal Lobule	R	4.71	32, -50, 42
Supramarginal Gyrus	L	4.17	-42, -44, 44
Thalamus	L	3.95	-12, -10, 4
	R	3.37	20, -24, 8

Table S6. Abstract category structure. Each row represents a unique stimulus (i.e., beetle). The four values assigned to a stimulus denote the four stimulus dimensions (e.g., antenna, legs, etc.) assigned to a beetle. Each numeric value (1 or 2) represents a specific feature instantiation (e.g., red or green eyes). The first dimension represents the rule-relevant dimension. Most Hole A beetles have a 1 on the first dimension (e.g., red eyes) whereas most Hole B beetles have a 2 (e.g., green eyes). The first stimulus in each of the columns is therefore an exception.

Hole A Beetles	**Hole B Beetles**
2 2 2 2*	1 2 2 2*
1 1 1 2	2 1 1 2
1 1 2 1	2 1 2 1
1 2 1 1	2 2 1 1
*Exception item	
Recognition Test Foils	
1 1 1 1	
1 1 2 2	
1 2 1 2	
1 2 2 1	
2 2 2 2	
2 2 1 1	
2 1 2 1	
2 1 1 2	

Supplemental Methods: Model Formalism and Procedures

SUSTAIN

Input Stimulus Representation. An input stimulus is represented to SUSTAIN as a pattern of activation on input units that code the different stimulus features and possible values that these features can take. For each stimulus feature, *i* (e.g., a beetle's eye color), with *k* possible values (two in the present experiment; e.g., red or green eyes), there are *k* input units. Input units are set to one if the unit represents the feature value or zero otherwise. The entire stimulus is represented by $I^{pos_{ik}}$, with *i* indicating the stimulus feature and *k* indicating the value for feature *i*. "Pos" indicates that the stimulus is represented as a point in a multidimensional space.

The distance μ_{ij} between the *i*th stimulus feature and cluster *j*'s position along the *i*th feature is

$$\mu_{ij} = 1/2 \sum_{k=1}^{v_i} | I^{pos_{ik}} - H_j^{pos_{ik}} | \qquad \text{(S1)}$$

where v_i is the number of possible values that the *i*th stimulus feature can take and $H_j^{pos_{ik}}$ is cluster *j*'s position on the *i*th feature for value *k*. Distance μ_{ij} is always between 0 and 1, inclusive.

Response Selection. After encoding, each cluster is activated based on the similarity of the cluster to the input stimulus. Cluster activation is given by:

$$H_j^{act} = \frac{\sum_{i=1}^{n_a} (\lambda_i)^\gamma e^{-\lambda_i \mu_{ij}}}{\sum_{i=1}^{n_a} (\lambda_i)^\gamma} \quad \text{(S2)}$$

where H_j^{act} is cluster *j*'s activation, n_a is the number of stimulus features, and λ_i is the receptive field tuning (which controls the model's attention to a given feature) for feature *i*, and *r* is the attentional parameter (constrained to be non-negative). λ_i is set to 1 at the start of learning.

Clusters enter into a competition to respond to an input stimulus through mutual inhibition. The final output of each cluster *j* is H_j^{out}, which is given by:

$$H_j^{out} = \frac{(H_j^{act})^\beta}{\sum_{i=1}^{n_c} (H_j^{act})^\beta} H_j^{act} \quad \text{(S3)}$$

where n_c is the current number of clusters and β is a lateral inhibition parameter (constrained to be non-negative) that controls the level of cluster competition.

Only the cluster with the largest output value is allowed to pass its output, H_j^{out}, across its connections to the output layer. The cluster that wins the competition, H_m, passes its output to the *k* output units of the unknown feature dimension *z*

$$C_{zk}^{out} = w_{m,zk} H_m^{out} \quad \text{(S4)}$$

where C_{zk}^{out} is the output of the unit representing the *k*th feature value of the *z*th feature, and $w_{m,zk}$ is the weight from the winning cluster, H_m, to the output unit C_{zk}. In the simulations present here where only the category label is queried, z always stands for the category label, and C_{zk}^{out} is calculated for each of the *k* (*k*=2) values that the category label can take (Hole A or B).

The probability of making a response *k* for a queried dimension, *z*, on a given trial is:

$$P(k) = \frac{e^{(dC_{zk}^{out})}}{\sum_{j=1}^{v_z} e^{(dC_{zk}^{out})}} \quad \text{(S5)}$$

Cluster recruitment. In the present simulation, SUSTAIN was initialized with two clusters, one for each category, that represent the modal stimulus in the category, excluding exceptions. This initialization reflects the cuing procedure used in the study, which alerted subjects to the rule-relevant dimension. Other clusters are recruited in SUSTAIN in response to surprising stimulus. Similar to recent uses of SUSTAIN (Love & Gureckis, 2007), we included a hippocampal function parameter, τ_h (constrained to be between 0 and 1) that probabilistically determines whether an error will lead to new cluster recruitment. If SUSTAIN makes a prediction error, and τ_h exceeds q, where q is a randomly generated value (between 0 and 1), a new cluster is recruited. If the value of τ_h is 1, all errors lead to new cluster recruitment, and the model is equivalent to SUSTAIN's original formulation (i.e., Love, Medin, & Gureckis, 2007).

Learning. The learning rules determine how the clusters are updated. Only the winning clusters are updated. If a new cluster is recruited on a trial, it will be the winner. Otherwise, the cluster that is most similar to the current stimulus will be the winner. The winning cluster H_m, will have its position adjusted by:

$$\Delta H_m^{pos_{ik}} = \eta(I^{pos_{ik}} - H_m^{pos_{ik}}) \qquad \text{(S6)}$$

where η is the learning rate parameter. The result of the updating is that the winning cluster moves toward the current stimulus. Over the course of learning, each cluster will tend toward the center of its members.

Receptive field tunings are updated according to

$$\Delta\lambda i = \eta e^{-\lambda_i \mu_{im}}(1 - \lambda_i \mu_{im}) \qquad \text{(S7)}$$

where m indexes the winning cluster.

The weights from the winning cluster to the output units are adjusted by the one layer delta learning rule (Rumelhart, Hinton & Williams, 1986).

$$\Delta w_{m,zk} = \eta(t_{zk} - C_{zk}^{out}) H_m^{out} \qquad \text{(S8)}$$

Simulations. For the present simulation, items were presented to SUSTAIN using the same order and randomization methods as for human subjects. To reflect the rule cuing procedure, the model was initialized with a cluster for each category that represented the average of the rule-following items within the category, and the attention weight on the rule-relevant dimension was initialized and fixed at 5.1 (cf. Love & Gureckis, 2007). The free parameters, γ, β, η, d, and τ_h, were fit to the average learning curve using standard optimization techniques. Obtained values were: γ = 1.259, β = 0.702, η = 0.415, d = 25.264, τ_h = 0.174.

The recognition strength measure R was calculated by summing the output H_j^{out} for all clusters:

$$R = \sum_{j=1}^{n_c} H_j^{out} \quad \text{(S9)}$$

The error correction measure was given by the summing absolute value of the difference between the winning cluster's output on unit k, for the queried feature dimension, z, and the item's actual position on this dimension.

$$E = \sum_{k=1}^{v_z} | C_{zk}^{out} - I^{pos_{zk}} | \quad \text{(S10)}$$

where v_z, again, equals the number of output units for feature dimension z.

Trial-by-trial values for recognition strength and error correction measures were averaged over 1000 simulations.

ALCOVE

Input Stimulus Representation. A stimulus is represented to ALCOVE as a pattern of activation a along input nodes that code the stimulus' values v for each feature dimension i. An entire stimulus is represented by the column vector a^{in}. Input nodes are gated by dimensional attention strength α_i that reflects the relevance of each feature dimension for the task.

Input stimuli activate each hidden node according to the psychological similarity of the input stimulus to the hidden node, where hidden nodes are each of the previously observed exemplars. The activation of the jth hidden node is:

$$a_j^{hid} = e^{-c(\sum_j \alpha_i |hji - a_i^{in}|^r)^{q/r}} \quad \text{(S11)}$$

where c a constant ($c > 0$) that determines the specificity of the hidden node, and where r and q are constants that determine the psychological distance metric and similarity gradient, respectively. Both r and q are set to 1 in the present applications, which corresponds to city-block distance metric with an exponential similarity gradient.

Response Selection. Hidden nodes are connected to k output nodes that reflect the available response categories (i.e., hole A or hole B) by a weight w_{kj} that gives the strength of the association between hidden node j and output node k. Output node activation is given by

$$a_k^{out} = \sum_{\substack{hid \\ j}} w_{kj} a_j^{hid} \quad \text{(S12)}$$

The probability of classifying a given stimulus into category k on a trial is given by

$$\Pr(K) = \frac{e^{\phi a_k^{out}}}{\sum_{\substack{out \\ k}} e^{\phi a_k^{out}}} \qquad \text{(S13)}$$

where is ϕ a mapping constant.

Learning. The dimensional attention strengths α_i and association weights w_{kj} are learned via gradient descent on sum squared error. The error generated by the model on a given trial is given by

$$E = 1/2 \sum_{\substack{out \\ k}} (t_k - a_k^{out})^2 \qquad \text{(S14a)}$$

where t_k are teacher values that code the feedback given to ALCOVE such that

$$tk = \left\{ \begin{array}{l} \max(+1, a_k^{out}) \text{ if the stimulus is in category K} \\ \min(-1, a_k^{out}) \text{ if the stimulus is not in category K} \end{array} \right\} \qquad \text{(S14b)}$$

The dimensional attention strengths and association weights are updated on each trial according to the learning rules

$$\Delta w_{kj}^{out} = \lambda_w (t_k - a_k^{out}) a_j^{hid} \qquad \text{(S15)}$$

$$\Delta \alpha = -\lambda_\alpha \sum_{\substack{hid \\ j}} \left[\sum_{\substack{out \\ k}} (t_k - a_k^{out}) w_{kj} \right] a_j^{hid} c \, | h_{ji} - a_i^{in} | \qquad \text{(S16)}$$

where λw and $\lambda \alpha$ are learning rate constants that are constrained to be greater than 0.

Simulations. ALCOVE was trained using the same procedure as the SUSTAIN simulations, with obtained values for the free parameters: $\lambda\alpha = 0.01$, $\lambda w = 0.40$, $c = 0.66$, $\phi = 3.125$.

Recognition strength is computed in ALCOVE as familiarity or similarity of an item to all exemplars stored in memory (i.e., the sum of the hidden node activations).

$$\sum_{\substack{hid \\ j}} a_j^{hid} \qquad \text{(S17)}$$

Error is computed for each trial as in S14a.