NeuroImage

https://doi.org/10.1016/j.neuroimage.2018.11.022

Brain reading and behavioral methods provide complementary perspectives on the representation of concepts

Andrew James Bauer[1] and Marcel Adam Just[2]

[1]*Department of Psychology, University of Toronto, Toronto, Ontario, Canada*
[2]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA.*

Keywords:
Neural representation
Concept representation
Semantic memory
fMRI
MVPA

ABSTRACT

The advent of brain reading techniques has enabled new approaches to the study of concept representation, based on the analysis of multivoxel activation patterns evoked by the contemplation of individual concepts such as animal concepts. The present fMRI study characterized the representation of 30 animal concepts. Dimensionality reduction of the multivoxel activation patterns underlying the individual animal concepts indicated that the semantic building blocks of the brain's representations of the animals corresponded to intrinsic animal properties (e.g. *fierceness*, *intelligence*, *size*). These findings were compared to behavioral studies of concept representation, which have typically collected pairwise similarity ratings between two concepts (e.g. Henley, 1969). Behavioral similarity judgments, by contrast, indicated that the animals were organized into taxonomically defined groups (e.g. *canine*, *feline*, *equine*). The difference in the results between the brain reading and behavioral approaches might derive from differences in cognitive processing during judging similarities versus contemplating one animal at a time. Brain reading approaches may have an advantage in describing thoughts about an individual concept, owing to the ability to decode brain activation patterns elicited by the brief consideration of a single concept (e.g. word reading) without a complex cognitive or behavioral task (e.g. similarity judgments). On the other hand, some behavioral tasks may tend to evoke a concept from numerous perspectives, yielding a representation of the breadth and sophistication of the concept knowledge. These results suggest that neural and behavioral measures offer complementary perspectives that together characterize the content and structure of concept representations.

1. Introduction

The advent of brain reading techniques has enabled new approaches to the study of concept representation, based on analysis of the fine-grained brain activation patterns evoked by the contemplation of different concepts (O'Toole et al., 2007; Mitchell et al., 2008; Huth et al., 2016). Research on knowledge of objects and other concepts has traditionally employed methods that collect behavioral responses, such as dissimilarity ratings between pairs of objects or lists of their properties (Murphy, 2004). Brain reading research has added to the body of behavioral work in characterizing the knowledge of a range of concepts, including concrete objects such as animals and tools (e.g. Henley, 1969; Ruts et al., 2004; Just et al., 2010) and abstract entities such as emotions and academic physics concepts (Baucom et al., 2012; Kassam et al., 2013; Mason & Just, 2016). Moreover, the two approaches have been shown to yield convergent results regarding some of the content of the concepts (e.g. Weber et al., 2009; Connolly et al., 2012; Connolly et al., 2016). Thus a major focus of neurosemantic research has been to provide an additional level of analysis that strengthens the behavioral evidence regarding some of the content of the representation of different concepts.

Although previous research has described a similarity between brain reading and behavioral methods, the question of how concept representations might differ between the paradigms has received less attention. An even greater understanding of concept representation may be enabled by an assessment of the unique perspective afforded by brain reading methods (Barsalou, 2017). Semantic cognition refers to several fundamental facets of concept knowledge, including the breadth and organization of the knowledge stored in long-term memory,

and also the representation of individual concepts and the content evoked during their on-line retrieval (Lambon Ralph, 2014). Brain reading approaches may be particularly well suited for illuminating the nature of the representation of individual concepts, owing to their ability to decode patterns of neural activity elicited by the consideration of a concept in the absence of a complex task, such as pairwise dissimilarity judgments (Bauer & Just, in press). Moreover, rich patterns of activation can be evoked by brief periods of consideration of a concept (e.g. 1 to 3 seconds or less). Brain reading methods can minimize the scope of a concept's evoked representation, reducing semantic processing that might otherwise arise from spreading activation (Anderson, 1983) or explicit consideration of other concepts. These attributes of neural measures of concept representation thus grant brain reading approaches high construct validity in interrogating the semantic basis for thinking about individual concepts.

Studies of the neural representation of individual concepts have enjoyed much success in revealing the kinds of properties or semantic dimensions underlying a concept's evoked representation. For example, data reduction techniques such as factor analysis have revealed the principal dimensions underlying concepts of emotions, such as *valence* and *lust* (Kassam et al., 2013). fMRI and MEG studies have also used regression models to predict the spatially distributed activation patterns of object concepts, based on some of the objects' properties extracted from text corpora (Mitchell et al., 2008; Pereira et al., 2013) or generated by human participants (Chang et al., 2011; Sudre et al., 2012). Thus, neural measures of concept representation suggest that the basis for thinking about individual concepts corresponds to thoughts about some of the fundamental properties of a concept.

Previous research has revealed that concept representations are encoded by spatially diverse neuronal populations across multiple lobes of the brain (Huth et al., 2016; Bauer & Just, in press). There are several theories regarding the neural representation of concepts, one of which is that a concept is encoded in spatially distributed, modality-specific areas that are specialized to process a concept's sensorimotor and affective properties (Kiefer & Pulvermüller, 2012). A related theory, called the "hub-and-spoke" theory, further posits that a centralized hub integrates a concept's anatomically distributed representations of its properties (Patterson et al., 2007). A major commonality to these theories is that a concept's properties are neurally represented across different brain systems.

On the other hand, behavioral research has delineated both the various properties of concepts and also the structural form of domains of concepts. For example, multidimensional scaling techniques have been used to reveal some of the fundamental properties of concepts about odors, colors, and concrete objects such as animals (Henley, 1969; Berglund et al., 1972; Indow & Aoki, 1983; Shepard & Cooper, 1992). Behavioral measures have also been used to reveal the structural or organizational form that underlies the knowledge of a domain of concepts. For example, although lists of properties generated for individual animals have been reduced to a low-dimensional space defined by some of the animals' intrinsic properties, these same data are better modeled as a hierarchical tree structure that groups the animals into increasingly specific taxonomic groups (e.g. rodents and mammals, primates and rodents) (Kemp & Tenenbaum, 2008). Behavioral measures may therefore describe the breadth and structural form of concept knowledge stored in long-term memory, which subsumes the concepts' more fundamental properties. This may derive from the fact that many behavioral studies collect a large amount of responses. For example, studies that collect lists of properties of concepts have tended to instruct participants to thoroughly consider each concept (e.g. 10 properties listed per concept, Ruts et al., 2004). Moreover, in studies that collect dissimilarity ratings, each concept might be judged against a large number of other concepts, thereby evoking a concept from numerous perspectives (Henley, 1969).

Research that directly compares neural and behavioral measures of representations of the same concepts could thus lead to a greater understanding that distinguishes between on-line retrieval of individual concepts and the long-term, comprehensive knowledge of the concepts. This was the goal of the current fMRI study, which compared the neural representations of 30 animal concepts to pairwise dissimilarity ratings of the same animals collected in a previous behavioral study (Henley, 1969). Schematic depictions of the two tasks are shown in Figure 1. The behavioral data consisted of 29 dissimilarity ratings per animal, and the neural representations corresponded to the multivoxel activation patterns evoked by the participants' limited consideration of each individual animal. Spatial patterns of activation composed of individual voxels are thought to reflect neural population codes and thus to indicate representational content (Mur et al., 2009).

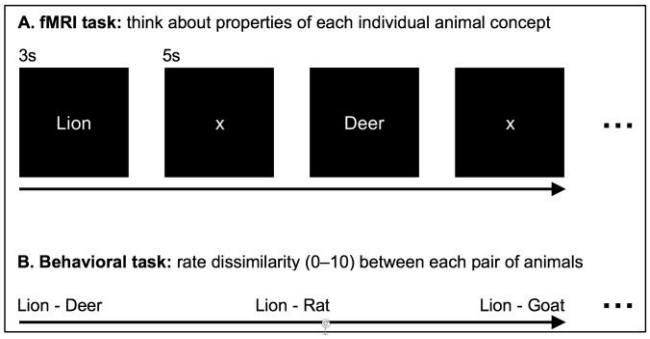

Figure 1. Schematics of the tasks performed in the (A) fMRI and (B) behavioral task. The representations of the same 30 animals evoked by the two tasks were compared. In the fMRI task, the participants thought about several properties of each individual animal. In the behavioral task, the participants rated the dissimilarity between each pair of animals, yielding 29 ratings per animal.

The approach taken here was to consider a distribution of brain areas that collectively contain the information about the animal concepts, consistent with previous research showing that concept representations are spatially distributed. Some studies have employed a searchlight procedure to locate small clusters of contiguous voxels whose activation patterns resemble behavioral dissimilarity ratings (Connolly et al., 2012; Connolly et al., 2016). While using for interrogating the representational content of regional clusters, the searchlight procedure cannot assess information encoded in a distributed network of brain areas (Haynes, 2015). The approach taken here was to better capture the totality of the spatially distributed neural representations without focusing the analysis only on regional activation that resembles the dissimilarity ratings. This approach enables an unbiased characterization of the neural representations to compare to the behavioral data.

Ever since the development of multidimensional scaling and clustering techniques, which model the structure of

similarity data, concept representations have traditionally been explored using behavioral similarity judgments (Borg & Groenen, 1997; Ruts et al., 2004). Only more recently have alternative similarity measures been employed, such as those based on statistical regularities between words in large text corpora. The current study therefore sought to illuminate how brain reading studies may offer a unique perspective on concept representation that complements the conclusions drawn from behavioral similarity ratings, which constitute the main source of information of how concepts are represented.

The current study compared both the content and the structural form of the animal representations evoked by the fMRI and dissimilarity rating paradigms. The 29 dissimilarity ratings per animal were hypothesized to implicitly encode a large number of properties of the animals, which should overlap with the more limited content expected in the individual animals' neural representations. Thus some degree of commonality was expected in the evoked content (Hypothesis 1), which was assessed using representational similarity analysis. This analysis enables relating together different measures of a putatively identical phenomenon by mapping the datasets to the same abstract space of inter-stimulus similarities which can then be compared (Shepard & Chipman, 1970; Kriegeskorte et al., 2008).

The structural form underlying the individual animals' neural representations was hypothesized to correspond to a low-dimensional space defined by intrinsic properties of the animals (Hypothesis 2A). This hypothesis draws from the multidimensional scaling analysis of the dissimilarity ratings originally reported in Henley (1969), which showed that the animal concept knowledge could be modeled by 3-dimensional space defined by the animals' *size*, *fierceness*, and *intelligence*. An exploratory factor analysis of the fMRI activation patterns was predicted to recover these three dimensions in addition to other intrinsic properties. Consistent with previous research on concept representation in the brain, the neural representations of the animal properties were predicted to correspond to activation patterns distributed across different sensory, motor, and affective areas. For example, the representation of the property *fierceness* might plausibly extend to areas underlying affective processing such as orbitofrontal cortex, as suggested by previous research that examined animal concept representations in the brain (Connolly et al., 2016).

On the other hand, a re-analysis of the ratings data from Henley (1969) was hypothesized to indicate a more complex underlying structure, specifically a partitioning of the animals into taxonomic groups (Hypothesis 2B). This hypothesis is consistent with more recent demonstrations of a taxonomic basis to lists of properties generated for various animals (Kemp & Tenenbaum, 2008). A factor analysis that requests several additional components should yield factors that divide the animals into taxonomic groups (e.g. a *feline* factor on which only members of the *feline* taxonomic family have high factor loadings). Furthermore, a clustering analysis should yield taxonomic groups that resemble a scientific taxonomic classification of the animals.

2. Material and methods

2.1. Participants

Twelve right-handed adults (four males, eight females; mean age of 23.9 years, ranging from 20 to 35) from Carnegie Mellon University and the Pittsburgh community participated and gave written informed consent approved by the Carnegie Mellon Institutional Review Board. Three additional participants' data were discarded due to falling asleep, and four other participants' data were excluded because of excessive head motion (greater than half the size of a voxel: 1.5mm total displacement in the *x* or *y* dimensions or 3mm in the *z* dimension). Four additional participants' data were discarded due to chance-level accuracy in multivoxel pattern classification of the animal concepts (classification features were the 120 most "stable" voxels selected from anywhere in the brain excluding the occipital lobe; more detail concerning classification is provided in section 2.4.4). This classification, which differed from the classification that tested the hypotheses, was used to check for systematicity in a participant's activation patterns regardless of its correspondence to the hypotheses.

2.2 Experimental paradigm and task

The stimuli in the fMRI task were 30 words each of which was the name of an animal, as shown in Table I. The 30 animals were the same 30 animals used in the previous behavioral study that collected pairwise dissimilarity ratings (Henley, 1969). During scanning, the 30 words were presented six times in six different random permutation orders (two permutation orders were presented during each of three scans). Each word was presented for 3s, followed by a 5s rest period, during which the participants were instructed to clear their minds and fixate on an "X" displayed in the center of the screen. There were nine total presentations of an "X" alone in the center of the screen, 24s each, distributed evenly throughout the three scans to provide a baseline measure for calculating percent signal change in the fMRI signal.

When a word was presented, the participants' task was to actively imagine and think about the properties of the animal to which the word referred. To promote their consideration of a consistent set of properties across the six presentations of an animal, the participants were asked to generate a set of three properties for each animal prior to the scanning session (for example, some properties generated for squirrel were "is small and nimble, eats nuts, climbs trees"). The participants were instructed to list the properties that came immediately to mind for each animal. Each participant was free to choose any properties for a given animal, and there was no attempt to impose consistency across participants in the choice of properties.

Table I. The 30 stimulus words (animal names) used in both the fMRI and behavioral tasks

Antelope	Donkey	Mouse
Bear	Elephant	Pig
Beaver	Fox	Rabbit
Camel	Giraffe	Raccoon
Cat	Goat	Rat
Chimpanzee	Gorilla	Sheep
Chipmunk	Horse	Squirrel
Cow	Leopard	Tiger
Deer	Lion	Wolf
Dog	Monkey	Zebra

2.3. fMRI scanning parameters and data preprocessing

Functional blood oxygen level-dependent (BOLD) images were acquired on a 3T Siemens Verio Scanner and 32-channel phased-array head coil (Siemens Medical Solutions, Erlangen, Germany) at the Scientific Imaging and Brain Research Center of Carnegie Mellon University using a gradient echo EPI sequence with TR = 1000ms, TE = 25ms, and a 60° flip angle. Twenty 5mm-thick AC-PC-aligned slices were imaged with a gap of 1mm between slices, in an interleaved spatial order starting at the bottom. The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5mm in-plane resolution.

Data preprocessing was performed with the Statistical Parametric Mapping software (SPM8, Wellcome Department of Cognitive Neurology, London, UK). Images were corrected for slice acquisition timing, motion, and linear trend; temporally smoothed with a high-pass filter using a 190s cutoff; and normalized to the Montreal Neurological Institute (MNI) template without changing voxel size (3.125 × 3.125 × 6mm).

The percent signal change relative to the baseline condition was computed at each gray matter voxel for each stimulus presentation, using SPM8. The main input measure for the subsequent analyses consisted of the mean of the four brain images acquired within a 4s window, offset 5s from the stimulus onset (to account for the delay in hemodynamic response). The intensities of the voxels in this mean image for each stimulus presentation were then normalized (mean = 0, SD = 1).

2.4. Data analysis

2.4.1. Overview

Representational similarity analysis was used to assess the commonality of the content encoded in the animal representations evoked by the brain reading task and behavioral dissimilarity judgements. Additionally, each dataset from these two tasks was separately submitted to exploratory factor analysis and clustering analysis to ascertain its underlying structural form.

2.4.2. Assessing the representational similarity between the brain activation patterns and dissimilarity ratings

To compare the two datasets, the fMRI activation patterns (described in section 2.4.3) were first converted into a vector of pairwise neural dissimilarities, and then correlated with the vector of pairwise dissimilarity ratings. (The activation data used for the representational similarity analysis corresponded to the same 255 stable voxels submitted to the factor analysis.) First, the vector of the pairwise neural dissimilarities was formed by computing the correlation distance (i.e. 1 – the Pearson correlation) between the activation patterns of each pair of animals, averaged across the six presentations of each animal. (Correlation distance has been shown to provide better accounts than other dissimilarity metrics, such as Euclidean distance; see Kriegeskorte et al., 2008.) This resulted in a vector of correlation distances, or dissimilarities, for the 435 possible pairs of animals. This vector of neural dissimilarities, averaged across participants, was then correlated with the vector of behavioral dissimilarities, which yielded an index of the commonality of the information carried in the two datasets.

2.4.3. Factor analysis of the brain activation patterns

The brain activation associated with the 30 animal concepts was factored into different components shared across participants using a two-level exploratory factor analysis, as described in previous studies (e.g. Just et al., 2010). The factor analysis was based on principal axis factoring with varimax rotation, implemented in MATLAB 7 (Mathworks, MA) using the same algorithm as the SAS factor procedure (www.sas.com).

The goal of the first-level factor analysis was to find the participant-specific distributed brain networks involved in the representation of the animal concepts. The first-level factor analysis was performed separately for each participant, resulting in 10 first-level factors. (The number of first-level factors was fixed at 10, which was the modal number of factors for all participants based on the Kaiser criterion.) These factors were characterized by their vector of scores for the 30 animals and their associations with specific subsets of an initially selected set of 255 voxels. These voxels had the most "stable" activation profiles; the stability of a voxel was computed as the average pairwise correlation between its activation profiles (vector of its activation levels across the 30 animals) across the repetitions of the animals (Just et al., 2010). The choice of the particular number of voxels used as the input was motivated by similar analyses in previous studies (Just et al., 2010; Just et al., 2014; Mason & Just, 2016). For each participant, the 255 most stable voxels were selected from five major brain areas, and the number of voxels selected from a brain area was proportional to that brain area's size: frontal lobe (100 voxels), temporal lobe (45), fusiform gyrus (15), parietal lobe (55), and occipital lobe (40). (The selection of voxels from different brain areas was motivated by the assumption that a semantic factor would be composed of a large-scale cortical network with representation in multiple brain areas.) The fusiform area was separated from the other areas because of its prominence in previous studies of object representation. The voxels were assigned to anatomical areas using Anatomical Automatic Labeling (AAL) (Tzourio-Mazoyer et al., 2002). Before the selection of stable voxels, occipital voxels were removed whose activation levels correlated with the character lengths of the stimulus words.

A second-level factor analysis was then run to identify factors that were common across the participants. The number of factors in this group-level factor analysis was limited to 6, beyond which they were not easily interpretable. The group-level factors were characterized by their vector of scores for the 30 animals and their associations with specific subsets of the first-level factors, and, through these associations, to subsets of originally selected voxels.

Note that a large number of voxels (255) was selected to better capture the majority of the activation data that jointly constitute the neural representation of the animal concepts. (A number of stable voxels close to 250 was selected such that the number of voxels chosen from a given major lobe was proportional to the size of that brain area, which resulted in a total of 255 voxels.) Also, the stable voxels were selected in proportion to each major brain area, rather than from anywhere in the brain, in order to prevent selection of voxels based solely on stability score. For example, because voxels from visual brain regions tend to have higher stability scores, a disproportionate number of occipital voxels would result from an agnostic selection criterion based only on stability. To compare the results to this agnostic selection method, 250, 500, and 1000 stable voxels were selected from anywhere in the brain (excluding visual voxels that correlated with length of the word stimuli). Representational similarity analysis was used to assess whether the information in these additional sets of activation patterns were similar to the activation data submitted to the main analyses.

2.4.4. Testing the interpretation of the factors underlying the neural representations

The factors that emerged were initially interpreted by observing which animals had the highest and lowest factor scores for a given factor, and by observing the brain locations of the factors. As detailed in the results, the semantic labels initially assigned to the factors were tested by comparing the animals' scores on a particular factor to two independent measures: independent behavioral ratings of the animals with respect to that factor, and the animal names' co-occurrence with words that describe that factor in a large corpus of text (i.e. latent semantic analysis).

Multivoxel pattern classification was also performed to determine whether an animal could be identified based on its activation pattern composed of the brain locations associated with the factors. Classification proceeded through three stages: algorithmic selection of a set of voxels (features) to be used for classification; training of a classifier on a subset of the data; and testing of the classifier on the remaining subset of the data. The training and testing used cross-validation procedures that iterated through all possible partitionings of the data into training and test sets, always keeping the training and test sets separate. The classification was performed with a support vector machines classifier that used 120 "stable" voxels (described above), where 20 of the most stable voxels were drawn from the clusters of voxels associated with each of the six factors. (Set sizes of 90 and 150 voxels, where 15 and 25 voxels were drawn from the clusters of voxels associated with each factor, resulted in similar classification accuracies, reported in section 3.2.2.) For each partitioning into training and test data, the voxel selection criterion was applied to the training set and the classifier was trained to associate an activation pattern to each of the 30 animals. Four (out of the six) repetitions of each animal were used for training and the mean of the remaining two repetitions was used for testing, resulting in 15 total partitionings into training and test data. The activation values of the voxels were normalized (mean = 0, SD = 1) across all the animals, separately for the training and test sets, to correct for possible drift in the signal across the six repetitions. Classification rank accuracy (referred to as accuracy) was the percentile rank of the correct word in the classifier's ranked output (Mitchell et al., 2004).

2.4.5. Factor analysis of the dissimilarity ratings

A single-level factor analysis was performed on the mean dissimilarity ratings (across participants) in Henley (1969). (The data for the individual participants were not available.) The number of factors requested was set to 6, as in the factor analysis of the fMRI data. This number was greater than the three dimensions visualized in the original multidimensional scaling analysis in Henley (1969).

In Henley (1969), 21 participants rated the dissimilarity between each pair of animals on a 0–10 scale, for a total 435 pairs (all possible pairs of the 30 animals). The ratings were repeated in a second session in which the positions of the animal names presented in each pair were inverted, and the ratings from both sessions were averaged together. Three participants' data with the greatest deviation from the sample had been removed and the remaining 18 participants' data were averaged together.

2.4.6. Testing the interpretation of the factors underlying the dissimilarity ratings

To test the hypothesis that the factors underlying the dissimilarity ratings correspond to taxonomic groups of animals (e.g. *rodent*), the groups of animals indicated in the factors were compared to a scientifically-defined taxonomic classification of the animals, as detailed in the results. *K*-means clustering of the dissimilarity ratings was used a data-driven method to separate the animals into the taxonomic groups suggested by the factors. The degree of agreement in the animals' group assignments between the dissimilarity ratings and taxonomic classification was quantified using the Wallace coefficient of congruence (Wallace, 1983). This agreement was then statistically assessed against chance using the null hypothesis testing procedure in http://www.comparingpartitions.info (Pinto et al., 2008), as detailed in the results.

3. Results

3.1. Modest commonality in the content of the animal representations evoked by the fMRI and behavioral tasks

There was some commonality in the content of the animal representations evoked by the two paradigms. The pairwise neural dissimilarities among the animal concepts (obtained by using representational similarity analysis) were modestly but statistically reliably correlated with the behavioral dissimilarity ratings: $r(433) = 0.12$, $p < 0.05$. This finding of representational commonality across the two types of measures was expected (Hypothesis 1), although the representational formats of the animal concepts were hypothesized to differ. As detailed below, in the fMRI measure of thinking of the individual animals, the principal dimensions underlying their neural representations corresponded to intrinsic properties of the animals (e.g. *size*, *intelligence*, *habitat*). On the other hand, the factors underlying the pairwise dissimilarity ratings corresponded to taxonomic groups of the animals (e.g. *rodent*, *feline*, *canine*).

3.2. Neural representations of the animals, considered individually, were underpinned by intrinsic properties of the animals

The six factors underlying the multivoxel brain activation patterns of the animal concepts corresponded to the animals' *fierceness* (two factors), *intelligence*, *habitat*, *farm-relatedness*, and *size*. The factor analysis was performed across participants, which points to the generality of the findings.

Figure 2 shows how the factor scores ordered the animals along each factor. For example, for the first *fierceness* factor (factor 1), the highest scores occurred for *bear* and *elephant* and some of the lowest scores occurred for *donkey, mouse,* and *sheep*. Here fierceness seems to refer to an animal's predacity or how threatening it is due to its size. The *intelligence* factor accorded its highest scores to *chimpanzee* and *dog*, and its lowest scores to *mouse* and *beaver*. Another factor appeared to reflect the *degree of enclosure of habitat* (e.g. den or burrow versus open field), and it assigned its highest scores to animals such as *lion, beaver,* and *bear* and some if its lowest scores to *horse, antelope,* and *camel* (animals that graze on open land). A fourth factor seemed to refer to another kind of *fierceness*, and it ranked *wolf* and *fox* highest and *goat* lowest. A fifth factor (*farm-relatedness*) favored many farm animals, assigning some of its highest scores to *horse, cow,* and *goat*. A sixth factor encoded *size*, according some of its highest scores to *elephant, zebra,* and *giraffe*, and its

lowest scores to *squirrel* and *chipmunk*. The percentage of variance accounted for by each of the six factors was *fierceness*: 7.3; *intelligence*: 6.3; *enclosure of habitat*: 5.7; second kind of *fierceness*: 5.4; *farm-relatedness*: 5.0; and *size*: 4.9.

Note that the results of the factor analysis are unlikely to have critically depended on the voxel selection method employed, namely the selection of voxels proportional to the size of each major brain area. Representational similarity analysis was used to assess whether the information encoded in the voxels submitted to the factor analysis was similar to the content encoded in 250, 500, and 1000 voxels selected from anywhere in the brain. The vector of pairwise neural dissimilarities (averaged across participants) computed from the voxels used in the main analyses was statistically reliably correlated with the dissimilarities computed from each set of voxels for comparison: $\underline{r}(433) = 0.18$, $p < 0.001$ (compared to 250 voxels); $\underline{r} = 0.22$, $p < 0.001$ (500 voxels); and $\underline{r} = 0.20$, $p < 0.001$ (1000 voxels).

Since each animal had a score for each of the factors, an animal concept's neural representation was a composition of these six factors. For example, *mouse* ranked high in *enclosure of habitat* and low in *intelligence*. The interpretations of the factors that are presented here are supported by the converging evidence below.

3.2.1. Brain areas associated with each factor

Each of the six factors was associated with multiple brain locations, distributed across multiple lobes. The clusters of voxels associated with each factor (identified by their high factor loadings) are shown in Figure 3. The *fierceness* factor (factor 1) showed high factor loadings for a cluster of voxels located in left orbitofrontal cortex which has been implicated in affective evaluation (e.g. reward and punishment in decision making) (Montague & Berns, 2002; Mitchell, 2009). A cluster of voxels corresponding to the second *fierceness* factor (factor 4) was also located in this brain area, and there were additional clusters in medial prefrontal cortex and left temporoparietal junction, which have also been implicated in social cognition (e.g. reasoning about another's beliefs or intent) (Samson et al., 2004; Mitchell, 2009). This network of regions implicated in social cognition has also been identified in a previous study that searched for multivoxel representations of an animal's predacity (Connolly et al., 2016). Connolly et al. (2016) found activation related to animal fierceness in superior temporal sulcus and other social cognition areas, and conclude that perception of threat extends to thoughts about non-human animals. These previous findings, along with the ordering of the animals along these two factors, provide support for the *fierceness* interpretations of factors 1 and 4.

The brain locations corresponding to *intelligence* (factor 2) were widely distributed in the brain. Among these locations were left inferior frontal gyrus and left middle temporal gyrus, which have been widely implicated in language (Price, 2010) and have been shown to underlie the representation of abstract concepts (e.g. *dogma*) more than concrete concepts (e.g. *screwdriver*) (Binder et al., 2005; Wang et al., 2010; Wang et al., 2013). There were also clusters of voxels associated with *intelligence* in bilateral superior angular gyrus/intraparietal sulcus, a heteromodal hub thought to organize information from different modalities and thus to construct abstract concepts (Binder & Desai, 2011; Bonner et al., 2013). Thoughts about intelligent animals could plausibly be multifaceted or abstract, referring to their complex methods of communication or social interactions.

The brain locations associated with *enclosure of habitat* (factor 3) notably included bilateral superior retrosplenial cortex/precuneus. These areas have been shown to activate to information about scenes and dwellings (Just et al., 2010; Bauer & Just, 2015), and are thought to be important for spatial navigation (Epstein & Higgins, 2007; Vann et al., 2009). Thus these areas could underlie thoughts about the spatial configuration or topography of an animal's habitat. There were also clusters of voxels associated with this factor in medial prefrontal cortex. Previous research suggests that medial prefrontal cortex and precuneus, where activation was also observed for this factor, constitute a circuit that underlies perspective-taking and the representation of the self in relation to the external world (Gusnard & Raichle, 2001; Cavanna & Trimble, 2006). Thus, the presence of activation in these clusters may indicate that thoughts about an animal include taking the perspective of an animal in relation to its habitat.

The factor *farm-relatedness* (factor 5) included clusters in bilateral inferior gyrus/precentral gyrus, in particular the mouth and face areas of the motor cortex. These areas have been shown to underlie food and eating concepts, for example concepts of vegetables and eating utensils, and concepts of mouth actions such as chewing (Hauk et al., 2004; Just et al., 2010; Carota et al., 2012; Bauer & Just, 2015; Carota et al., 2017). The information represented in these areas might therefore refer to thoughts of consumption of farm animal meat or other farm animal products.

Chief among the brain locations associated with *size* (factor 6) were bilateral parahippocampal gyrus, which has been implicated in the representation of object size (Konkle & Oliva, 2012), and bilateral intraparietal sulcus, important to visuospatial working memory (Todd & Marois, 2004). Additional clusters of voxels associated with *size* were located in visual regions such as lateral occipital complex (namely bilateral middle and superior occipital gyrus).

The attributions accorded to the brain locations (and hence factors) above are instances of reverse inference, which are complemented by the orderings of the animals along the factors.

3.2.2. Animal identification accuracy based on the factors

It was possible to identify which animal concept a person was thinking about with accuracies above chance level by training a classifier on a subset of that person's brain activation patterns (four out of six presentations) and then making the identification over an independent dataset (the mean of the remaining two presentations). The classification was performed using voxels drawn from the clusters of voxels discovered to underlie each of the six factors. The resulting mean classification rank accuracy (averaged across participants) was 0.74, with a range of 0.60 to 0.86. The accuracies for all the individual participants were above chance level, as determined by permutation testing (the $p < 0.001$ probability threshold for a rank accuracy being greater than chance level is 0.54). (Furthermore, set sizes of 90 and 150 voxels resulted in similar mean classification accuracies, namely 0.73 and 0.75.)

The results of the classification analysis indicate that the pattern information represented in the brain locations associated with the factors systematically differed for the animal concepts. These results are expected if the uncovered factors correspond to different semantic dimensions of animal concepts that jointly define the animals.

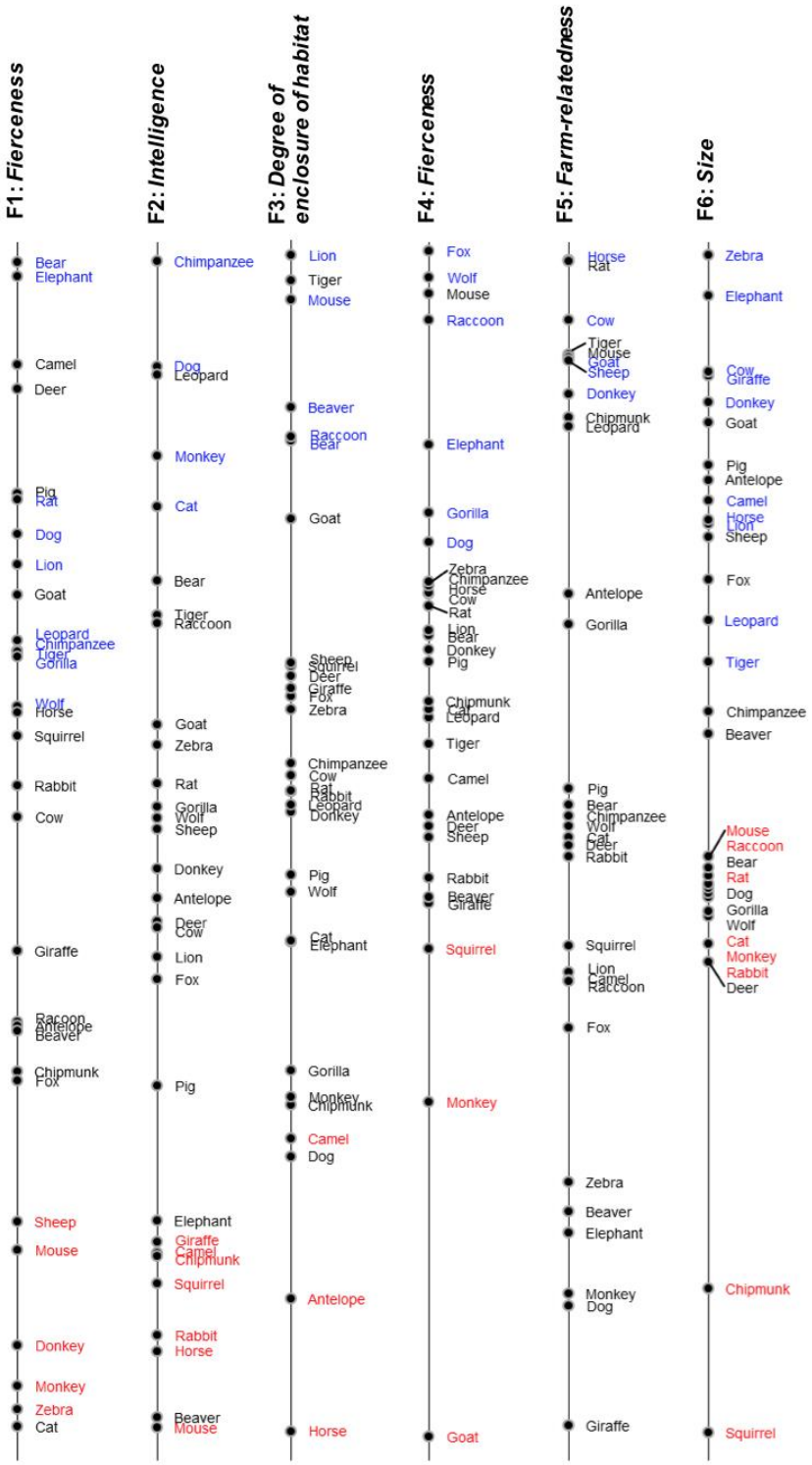

Figure 2. The factors underlying the neural representations of the animals corresponded to intrinsic properties of the animals. For each factor shown above, the 30 animals are arranged on a line according to their factor scores. In each factor there were two sets of animals with high face validity for the factor label: animals that had high factor scores (colored in blue) or low scores (red). For *farm-relatedness* (factor 5) only the animals with high factor scores could be interpreted as having high face validity for that factor.

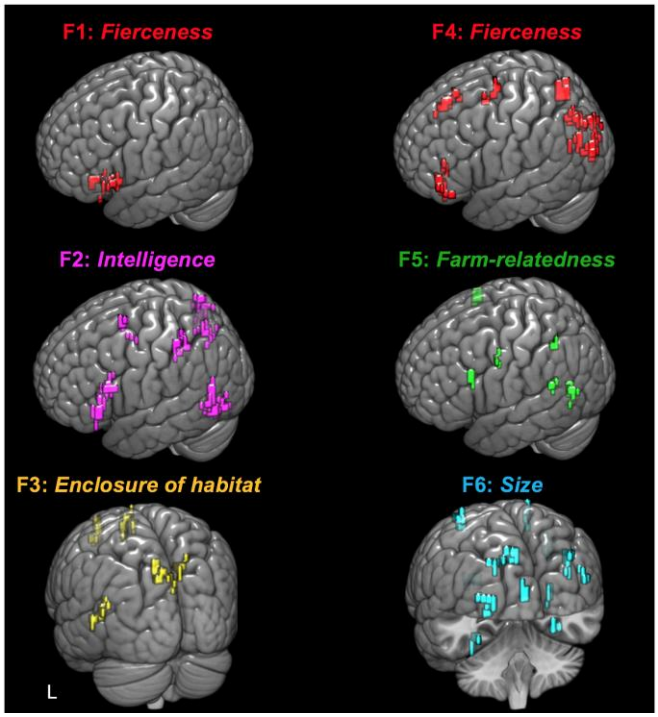

Figure 3. The brain locations associated with each of the factors underlying the neural representations of the animals. In the fMRI measure of thinking of the individual animal concepts, the neural representations were underpinned by intrinsic properties of the animals. Shown above are the clusters of voxels associated with each of the factors that emerged from a cross-participant exploratory factor analysis, which is the union of the voxels from the individual participants found to correspond to the second-level factors. For factors 1–4, the voxels were clustered at a threshold of 10 voxels. For factors 5–6, the clustering threshold was set to 5 voxels, as these factors accounted for lower percentages of variance in the data. Rendering was performed on an MNI template brain using the 3D medical imaging software MRIcroGL (Rorden & Brett, 2000).
L: left; *MNI*: Montreal Neurological Institute

3.2.3. Testing the interpretation of the factors using independent behavioral ratings on animal properties

One test of the semantic labels assigned to the factors was obtained by comparing the factor scores to independent behavioral ratings of the animals according to different animal properties. A previous study collected ratings of animals with respect to four properties, namely *intelligence*, *size*, *fierceness*, and *speed* (Holyoak & Mah, 1981). Three of these properties resembled four of the six uncovered factors (*intelligence*, *size*, and the two factors for *fierceness*). In Holyoak & Mah (1981), 25 participants rated the subjective magnitude of each property a 9-point scale, for each of the animals (28 animals were common between Holyoak & Mah, 1981, and the present fMRI study). For each property, the ratings of the 28 animals in common were correlated with the corresponding factor scores derived from the brain activation patterns. The mean ratings (averaged over participants) were correlated with the corresponding factor scores as follows: *intelligence*: $r(26) = 0.59$, $p < 0.001$; *size*: $r = 0.69$, $p < 0.001$; and *fierceness* (factor 1): $r = 0.36$, $p = 0.06$ (marginally significant). The other factor for *fierceness* (factor 4) was not correlated with the ratings, suggesting that this factor has a more limited association with *fierceness*: ($r = 0.24$, $p = 0.23$). The low correlations with the behavioral ratings of *fierceness* suggests that a constellation of animal properties related to an animal's *fierceness* is represented in these two factors.

3.2.4. Testing the interpretation of the factors using latent semantic analysis

Three of the six factors did not correspond to any of the rated properties in Holyoak & Mah (1981). These factors were *enclosure of habitat*, *farm-relatedness*, and the second *fierceness* factor (factor 4). To test the interpretation of the semantic labels associated with these factors, latent semantic analysis (LSA, http://lsa.colorado.edu) was used, which represents a word's meaning by its statistical co-occurrence with other words in a large corpus of text (Deerwester et al., 1990; Landauer & Dumais, 1997). LSA was used to determine the distance between each of the 30 animal names and a string of one to three words (excluding the animal names) intended to describe a given factor. The string defined for *enclosure of habitat* was "hole den burrow"; for *farm-relatedness* it was "farm"; and for the second factor for *fierceness* it was "attack ferocious force." The resulting LSA-computed distances between each animal and the descriptive strings were correlated with the animals' corresponding factor scores derived from the brain activation patterns as follows: *enclosure of habitat*: $r(28) = 0.50$, $p < 0.01$; *farm-relatedness*: $r = 0.35$, $p = 0.06$ (marginally significant); and *fierceness* (factor 4): $r = 0.38$, $p < 0.05$, indicating that the second factor for *fierceness* retains some association with this attribute.

Note that the choice of LSA over WordNet was made to provide a linguistic interpretation of the factors to test the hypothesis that the factors correspond to animal properties. LSA

has been used by previous studies to provide a linguistic interpretation of factors (e.g. Just et al., 2010). The possibility of a taxonomic basis to the animals' neural representation was explored in the comparison of the *k*-means clusters of the neural representations with the scientifically-defined taxonomic groups, rather than through a comparison to WordNet's taxonomic structure (Miller, 1995). Although some studies have shown a correspondence between activation patterns of object concepts and taxonomic structure as modeled by WordNet, these studies considered objects across qualitatively different categories (e.g. animate and non-living categories) and thus demonstrate a significant effect of category boundary (e.g. Kriegeskorte, 2008; Fairhall & Caramazza, 2013; Carlson et al., 2014). By contrast, research that examined within-category neural similarity (e.g. only between animals) did not find a relationship with WordNet (Carlson et al., 2014).

3.2.5. Summary

The principal dimensions underlying the neural representations of the animal concepts corresponded to intrinsic properties of animals. The brain areas associated with a given factor have been shown in previous studies to be involved in perceptual, cognitive, or social processing related to that factor. Furthermore, two independent metrics obtained without brain imaging (i.e. behavioral ratings of animal properties, and corpus-based characterizations of word meaning) bore a substantial relation to the characterization obtained through factor analysis of the brain activation patterns. These results are consistent with the hypothesis that the fMRI data reveal the core properties of animal concepts evoked during contemplation of individual animals (Hypothesis 2A).

3.3. The factors underlying the dissimilarity ratings indicated taxonomic groups of the animals

The factors underlying the dissimilarity ratings ordered the animals according to the degree to which they fit into different taxonomic groups (taxa). These factors corresponded to *rodent*, *equine*, *feline*, *primate*, *canine*, and *bovine* taxa. (The *bovine* factor might also be interpreted as a *farm-relatedness* factor, although this is less likely given the taxonomic basis for the other factors.) Figure 4 shows how the factor loadings ordered the animals along each factor.

There were two clusters of animals in each factor: one cluster containing a few animals with high loadings on that factor, and one cluster with most or all of the remaining animals with low loadings on that factor. For example, in the *feline* factor (factor 3) there was a cluster of feline animals as well as a non-feline cluster. The *equine* factor (factor 2) had a cluster of equine-like animals with intermediate factor loadings, and some of these animals did not load high on any of the factors, namely *camel*, *giraffe*, and *elephant* (other ungulates, i.e. hoofed animals). The percentage of variance accounted for by each of the six factors was *rodent*: 15.3; *equine*: 13.9; *feline*: 10.4; *primate*: 7.2; *canine*: 5.6; and *bovine*: 5.4.

3.3.1. Testing the interpretation of the factors against a scientific taxonomic classification of the animals

To test the interpretation that the factors underlying the dissimilarity ratings correspond to different taxa, the groups of animals indicated in the factors were compared to a scientific taxonomic classification of the animals (as specified by the Integrated Taxonomic Information System, www.itis.gov). First, a data-driven method, *k*-means clustering, was applied to the dissimilarity ratings to separate the animals into the groups that were suggested by the factors. The number of requested clusters *k* was set to 7, which equaled the number of distinct taxonomic groups indicated in the factors (including a group of equine-like animals suggested in the *equine* factor, which did not load high in any factor). The scientifically-defined taxonomic groups of the animals were formed by grouping together animals that have the same classification at a taxonomic rank of *genus*, *family*, or *order* (the three ranks directly above *species*). There were nine such groups. Two of the 30 animals did not share their taxonomic classification with any of the other animals, and so constituted their own singleton groups.

The groups of animals from the clustering of the dissimilarity ratings and as defined from the scientific taxonomic classification are detailed in Table II. The degree of agreement in the

animals' group assignments between the dissimilarity ratings and taxonomic classification was quantified using the Wallace coefficient, which expresses this agreement on a continuous scale between 0–1 (Wallace, 1983). The degree of congruence between the classifications was high: $W_{taxonomy \rightarrow ratings}$ = 0.67, 95% CI [0.55, 0.79], meaning that if two animals are in the same group according to the taxonomic classification, they have a 67% chance of being in the same group assignment according to the clustering of the dissimilarity ratings. This degree of agreement was greater than the value expected by chance: the Wallace coefficient of independence $W_{i\ (taxonomy \rightarrow ratings)}$ did not fall within the $W_{taxonomy \rightarrow ratings}$ 95% confidence interval ($W_{i\ (taxonomy \rightarrow ratings)}$ = 0.13), hence the null hypothesis of independence between the classifications was rejected (Pinto et al., 2008). Furthermore, $W_{ratings \rightarrow taxonomy}$ = 0.54, 95% CI [0.34, 0.74], and the corresponding Wallace measure of independence $W_{i\ (ratings \rightarrow taxonomy)}$ did not fall within the $W_{ratings \rightarrow taxonomy}$ 95% confidence interval ($W_{i\ (ratings \rightarrow taxonomy)}$ = 0.10). Table III contains the contingency table that presents the numbers of animals that shared their group assignment between the two classifications.

By contrast, there was no statistically significant agreement between the scientifically-defined taxonomic groups of animals and *k*-means clusters of the fMRI data. (*K*-means clustering of the fMRI data yielded *k* = 7 clusters that were not easily interpretable.) $W_{fMRI \rightarrow taxonomy}$ = 0.12, 95% CI [0, 0.26], and the Wallace coefficient of independence $W_{i\ (fMRI \rightarrow taxonomy)}$ was within the $W_{fMRI \rightarrow taxonomy}$ 95% confidence interval ($W_{i\ (fMRI \rightarrow taxonomy)}$ = 0.10), hence the null hypothesis of independence between the classifications was not rejected. Furthermore, $W_{taxonomy \rightarrow fMRI}$ = 0.16, 95% CI [0, 0.32], and the Wallace coefficient of independence $W_{i\ (taxonomy \rightarrow fMRI)}$ also fell within the 95% confidence interval ($W_{i\ (taxonomy \rightarrow fMRI)}$ = 0.13). Table III contains the contingency table which presents the lack of congruence between the two classifications.

3.3.2. Summary

The factors of the behavioral dissimilarity ratings were consistent with a scientific taxonomic classification of the animals, indicating *rodent*, *feline*, *canine*, and other taxa. Taxonomic grouping reflects the degree to which animals are of the same kind based on multiple shared properties. These results agree with previous accounts of adults' taxonomic intuitions about animals (Kemp & Tenenbaum, 2008; Unger et al., 2016). The results are consistent with the hypothesis that the dissimilarity ratings reveal the domain-level structure of the knowledge of animals (Hypothesis 2B).

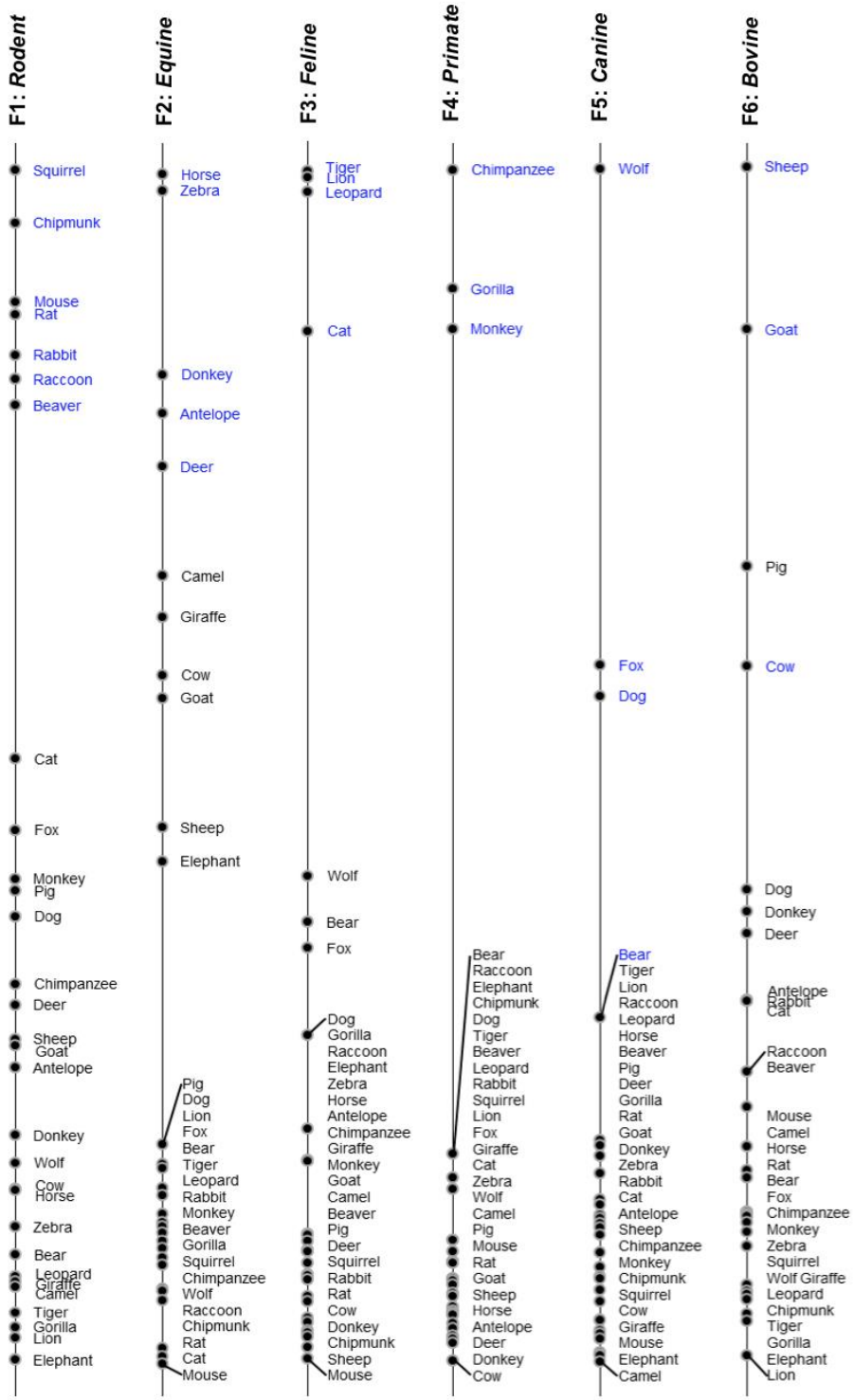

Figure 4. The factors underlying the dissimilarity ratings indicated a division of the animals into taxonomic groups. For each factor depicted above, the 30 animals are arranged on a line according to their loadings on that factor. In each factor there were two clusters of animals: one with few animals that had high face validity for the factor label and high loadings (colored in blue); and a second cluster with most or all of the remaining animals which loaded low. The *equine* factor (factor 2) had a cluster of equine-like animals with intermediate factor loadings, and some of these animals did not load high on any of the factors, namely *camel*, *giraffe*, and *elephant* (other ungulates, i.e. hoofed animals).

Table II. The clusters of the animals from the behavioral dissimilarity ratings were congruent with the scientific taxonomic groups of the same animals

Groups of animals from dissimilarity ratings

Rodents	*Equines*	*Felines*	*Primates*	*Canines*	*Bovids*	*Other ungulates (hoofed animals)*
Beaver	Donkey	Cat	Chimpanzee	Dog	Cow	Camel
Chipmunk	Horse	Leopard	Gorilla	Fox	Goat	Elephant
Mouse	Zebra	Lion	Monkey	Wolf	Sheep	Giraffe
Rat	Antelope*	Tiger			Pig	
Squirrel	Deer*	Bear*				
Rabbit*						
Raccoon*						

Groups of animals from scientific taxonomic classification

Rodents	*Equines*	*Felines*	*Primates*	*Canines and caniforms (dog-like carnivores)*	*Bovids*	*Other ungulates (hoofed animals)*	*Proboscideans (trunked animals)*	*Leporids (hares)*
Beaver	Donkey	Cat	Chimpanzee	Dog	Cow	Antelope	Elephant	Rabbit
Chipmunk	Horse	Leopard	Gorilla	Fox	Goat	Camel		
Mouse	Zebra	Lion	Monkey	Wolf	Sheep	Deer		
Rat		Tiger		Bear		Giraffe		
Squirrel				Raccoon		Pig		

Note. The degree of agreement in the 30 animals' group assignments between the *k*-means clusters from the dissimilarity ratings (top row) and the scientifically-defined taxonomic groups (www.itis.gov) (bottom row) was greater than the value expected from chance, as described in the text.
*A small number of animals in the behavioral groups were technically inconsistent with their group's taxonomic label according to the scientific classification.

Table III. Contingency tables demonstrating agreement of the scientific taxonomic classification with the behavioral dissimilarity ratings, but not with the fMRI data, in the animals' group assignments

		Groups of animals from scientific taxonomic classification								
		Rodents	***Equines***	***Felines***	***Primates***	***Canines and caniforms***	***Bovids***	***Other ungulates***	***Probos-cideans***	***Leporids***
Clusters of animals from dissimilarity ratings	***Rodents***	5				1				1
	Equines		3					2		
	Felines			4		1				
	Primates				3					
	Canines					3				
	Bovids						3	1		
	Other ungulates							2	1	

		Groups of animals from scientific taxonomic classification								
		Rodents	***Equines***	***Felines***	***Primates***	***Canines and caniforms***	***Bovids***	***Other ungulates***	***Probo-scideans***	***Leporids***
Clusters of animals from fMRI activation patterns	***Cluster 1***	1	1		1	1		1		
	Cluster 2	2						1		1
	Cluster 3			1	1	1	1	1		
	Cluster 4		1				2			
	Cluster 5	1	1			2				
	Cluster 6			2		1		1	1	
	Cluster 7	1		1	1			1		

Note. The contingency tables specify for each group of animals the numbers of those animals that belong to the groups of the other classification. Blank entries denote "0." *Top row*: Most or all of the animals in each cluster from the dissimilarity ratings were identically assigned according to the scientific taxonomic classification. *Bottom row*: The animals in the (unnamed) clusters from the fMRI data were randomly distributed among the groups of animals defined from the scientific taxonomic classification.

4. Discussion

This study demonstrates that brain reading and behavioral paradigms may provide complementary characterizations of the representations of concepts, using animal concepts as a case study of the comparison. The current study focused on comparing brain reading results to behavioral dissimilarity ratings, given that concept representations have traditionally been explored using dissimilarity judgments since the development of multidimensional scaling and clustering techniques, which model the structure of similarity data (Borg & Groenen, 1997; Ruts et al., 2004). Exploratory factor analysis was separately applied to multivoxel patterns of brain activation underlying animal concepts and to dissimilarity ratings between pairs of the same animals, revealing differences in their principal dimensions. Representational similarity analysis also indicated that there was some commonality in the representations revealed by the two paradigms. Taken together, the results suggest that fMRI measures of thinking of individual concepts reveal their intrinsic properties, whereas behavioral tasks such as dissimilarity ratings reveal the structure of the knowledge of a domain of concepts.

Concept knowledge refers to both the semantic content of individual concepts, and to the structure of that knowledge as manifested by the relationships among related concepts. A unified characterization of animal concept knowledge emerged from an analysis of complementary datasets, namely the neural representations of 30 animal concepts and dissimilarity ratingsbetween pairs of the same animals collected in a previous behavioral study (Henley, 1969). As hypothesized, the animal concepts' neural representations encoded a set of core properties of animals. On the other hand, the dissimilarity ratings encoded a scientifically-defined taxonomic organization of the concept knowledge of the animals. The 29 dissimilarity judgments made per animal might have been expected to reflect a richness of animal properties, including properties that provide contrasts among the animals, but the findings suggest that the main information underlying the behavioral dissimilarity ratings consists of inter-item taxonomic relatedness, thus indicating a domain-level structure of the concept knowledge about animals.

The results here complement previous research that related neural concept representations to behavioral dissimilarity ratings of the concepts. Using representational similarity analysis, previous studies found that dissimilarities of animal concepts based on multivoxel activation patterns in lateral occipital complex, but not in other brain areas, resembled dissimilarity ratings of the animals (Weber et al., 2009; Carlson et al., 2014; Connolly et al., 2012; Connolly et al., 2016). Lateral occipital complex has been shown to encode information about object geometry and other visual features (Grill-Spector et al., 2001). Notably, these studies investigated the neural representation of only a small number of animals (between 6-12), and three studies used animals from different classes (bugs, birds, reptiles, and mammals) (Connolly et al., 2012; Carlson et al., 2014; Connolly et al., 2016). Thus, information about body shape and other visual features may be sufficient to disambiguate a small number of animals from different animal classes, demonstrating an apparent taxonomic basis to their neural representations. These studies also found additional brain areas that contained information about the animal concepts, such as inferior frontal gyrus, precentral gyrus, middle temporal gyrus, and intraparietal sulcus. However, the dissimilarities among the neural representations of the animal concepts in these areas did not resemble the dissimilarities in the behavioral ratings. These areas, in addition to lateral occipital complex, were also found in the present study to underlie the neural representation of the animal concepts. Thus, although some brain areas may contain information about animals that appears to disambiguate them along taxonomic lines, a more comprehensive representation of animal concepts, distributed across many different brain areas, was found here to encode a limited number of intrinsic animal properties that do not constitute a scientific classification of animals more generally. The results reported here may possibly extend to representations of other types of concepts besides animals, such as tools and other manmade objects. That is, contemplation of a given individual concept may generally activate only a small number of intrinsic properties of the concept.

The differences in the principal dimensions underlying the representations highlight the perspective afforded by fMRI measures of concept representation. Brain reading methods enable a characterization of the representation of individual concepts. This is made by possible by the evocation of a large multivariate dataset of brain activation associated with a concept by even brief periods of consideration.

Brain reading paradigms might also be useful in revealing core, context-independent properties associated with an individual concept. A concept's neural representation can be elicited through minimal cognitive processing, thus circumventing a complex behavioral task that might introduce additional cognitive processes and influence the concept's properties that are evoked. The representation of a concept is thought to be variable and to depend on the evoking context (Yee & Thompson-Schill, 2016); thus "core properties" here refers to intrinsic properties that are most often evoked across different contexts. It may be interesting for future research to determine the extent to which spreading brain activation evokes additional properties of a concept, such as less critical properties and those not explicitly thought about (Collins & Loftus, 1975; Anderson, 1983).

Behavioral methods for determining the properties of a concept may also produce both intrinsic and taxonomic information, such as in a task requiring the listing of a concept's properties. These data may indicate the relationships among concepts and thus the structure of that knowledge, which can provide insight on how people reason using those concepts (Kemp & Tenenbaum, 2008). Dissimilarity ratings may be particularly well suited to focus on properties that distinguish concepts, indicating how they are organized in a larger representational space that subsumes many different concepts.

What has not yet been attempted is a brain reading study that uses a concept comparison task that resembles a dissimilarity rating task. Comparing brain reading and behavioral measures while keeping the task the same is an important consideration, given that the types of information retrieved about a concept have been shown to depend on the evoking task or context (Xu et al., 2018). Dissimilarity rating tasks can require large numbers of judgments, which are difficult to accommodate in an fMRI study. However, it is possible that such an fMRI task would reveal interesting additional details about taxonomically oriented neural representations, such as the properties of animal concepts that underlie the comparisons.

Another avenue of future research could be to explore how neural concept representations change as a function of the amount of time allotted to thinking about the concept stimuli. For example, studies that evoke event-related potentials present word stimuli tachistoscopically (on the order of 100ms stimuli

durations) (Hauk & Pulvermüller, 2004). One possibility is that minimal semantic processing gives rise to concept representations that are less enriched by semantic features (Bauer & Just, in press). Briefer presentations would also serve to further minimize confounds due to stimuli themselves. Future research could more precisely characterize the content and structure of concept representations that are activated during very brief periods of consideration, such as in everyday speech.

This study shows how a more complete characterization of concept knowledge emerges from complementary results from brain reading and behavioral methods such as dissimilarity ratings. Comparative research may enable a greater understanding of concept representation through future research or an integration of comparable existing datasets.

Acknowledgements

This research was supported by the National Institute of Mental Health grant MH029617, and the Office of Naval Research grant N00014-16-1-2694. We thank Charles Kemp for many helpful suggestions regarding the data analysis and writing of the paper.

References

Anderson, J. R. (1983). A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Barsalou, L. W. (2017). What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia*, *105*, 18–38. doi:10.1016/j.neuropsychologia.2017.04.011

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *NeuroImage*, *59*(1), 718–727. doi:10.1016/j.neuroimage.2011.07.037

Bauer, A. J., & Just, M. A. (2015). Monitoring the growth of the neural representations of new animal concepts. *Human Brain Mapping*, *36*(8), 3213–3226. doi:10.1002/hbm.22842

Bauer, A. J., & Just, M. A. *Neural representations of concepts*. To appear in G. de Zubicaray & N. Schiller (Eds.), *Oxford Handbook of Neurolinguistics*. Oxford: Oxford University Press.

Berglund, B., Berglund, V., Engen, T, & Ekman, G. (1972). Multidimensional analysis of twenty-one odors. Reports of the Psychological Laboratories, University of Stockholm (Report No. 345). Stockholm, Sweden: University of Stockholm.

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–917.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. doi:10.1016/j.tics.2011.10.001

Bonner, M. F., Peelle, J. E., Cook, P. A., & Grossman, M. (2013). Heteromodal conceptual processing in the angular gyrus. *NeuroImage*, *71*, 175–186. doi:10.1016/j.neuroimage.2013.01.006

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications.* New York: Springer-Verlag.

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2014). The emergence of semantic meaning in the ventral temporal pathway. *Journal of Cognitive Neuroscience*, *26*(1), 120–131. doi: 10.1162/jocn_a_00458

Carota, F., Moseley, R., & Pulvermüller, F. (2012). Body-part-specific representations of semantic noun categories. *Journal of Cognitive Neuroscience*, *24*(6), 1492–1509.

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex*, *27*(1), 294–309.

Cavanna, A. E. & Trimble, M. R. The precuneus: a review of its functional anatomy and behavioural correlates. (2006). *Brain*, *129*(3), 564–583. doi: 10.1093/brain/awl004

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., & Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience*, *32*(8), 2608–18. doi:10.1523/JNEUROSCI.5547-11.2012

Connolly, A. C., Sha, L., Guntupalli, J. S., Oosterhof, N., Halchenko, Y. O., Nastase, S. A., di Oleggio Castello, M. V., Abdi, H., Jobst, B. C., Gobbini, M. I., & Haxby, J. V. (2016). How the Human Brain Represents Perceived Dangerousness or "Predacity" of Animals. *Journal of Neuroscience*, *36*(19), 5373–5384. doi:10.1523/JNEUROSCI.3395-15.2016

Collins, A. M. & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review, 82*(6), 407–428.

Damarla, S. R., & Just, M. A. (2013). Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping*, *34*(10), 2624–34. doi:10.1002/hbm.22087

Davis, T., & Poldrack, R. A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral Cortex*, *24*(7), 1720–37. doi:10.1093/cercor/bht014

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048. doi:10.3758/BRM.40.4.1030

Epstein, R. A, & Higgins, J. S. (2007). Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cerebral Cortex*, *17*(7), 1680–1693. doi:10.1093/cercor/bhl079

Fairhall, S. L & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *The Journal of Neuroscience*, *33*(25), 10552–10558. doi: 10.1523/JNEUROSCI.0051-13.2013

Fisher, A. V., Godwin, K. E., Matlen, B. J., & Unger, L. (2015). Development of Category-Based Induction and Semantic Knowledge. *Child Development*, *86*(1), 48–62. doi:10.1111/cdev.12277

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*, 1409–1422.

Gusnard, D. A. & Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Review Neuroscience*, *2*, 685–694.

Hauk, O & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *5*(115), 1090—1103. doi: 10.1016/j.clinph.2003.12.020

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*, 257–280.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning & Verbal Behavior*, *8*, 176–184.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197–204. doi:10.3758/BF03202335

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. (2016). Natural speech reveals semantic maps that tile human cerebral cortex. *Nature*, *532*, 453–458.

Indow, T., & Aoki, N. (1983). Multidimensional mapping of 178 Munsell colors. *Color Research and Application*, *8*, 145–152.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, *5*(1), e8622. doi:10.1371/journal.pone.0008622

Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A., & Mitchell, T. M. (2014). Identifying Autism from Neural Representations of Social Interactions: Neurocognitive Markers of Autism. *PLoS ONE*, *9*(12), e113879. doi:10.1371/journal.pone.0113879

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PLoS One*, *8*(6), e66032. doi:10.1371/journal.pone.0066032

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*, *105*(31), 10687–10692.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. doi:10.1016/j.cortex.2011.04.006

Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124. doi:10.1016/j.neuron.2012.04.036

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4). doi:10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1634), 20120392. doi:10.1098/rstb.2012.0392

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Mason, R. A., & Just, M. A. (2016). Neural Representations of Physics Concepts. *Psychological Science*, *27*(6), 904–913.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., & Newman, S. D. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*, 145–175.

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A, & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. doi:10.1126/science.1152876

Mitchell, J. P. (2009). Social psychology as a natural kind. *Trends in Cognitive Sciences*, *13*(6), 246–51. doi:10.1016/j.tics.2009.03.008

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265–284. doi:10.1016/S0896-6273(02)00974-1

Mur, M., Bandettini, P. A, & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. doi:10.1093/scan/nsn044

Murphy, G. L. (2004). *The Big Book of Concepts*. Cambridge: MIT Press.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, *19*(11), 1735–1752. doi:10.1162/jocn.2007.19.11.1735

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews: Neuroscience*, *8*(12), 976–987. doi:10.1038/nrn2277

Pinto, F. R., Melo-Cristino, J., & Ramirez, M. (2008). A Confidence Interval for the Wallace Coefficient of Concordance and Its Application to Microbial Typing Methods. *PLoS ONE*, *3*(11). doi:10.1371/journal.pone.0003696

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*, 62–88. doi:10.1111/j.1749-6632.2010.05444.x

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology*, *12*(4), 191–200.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506–515. doi:10.3758/BF03195597

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience*, *7*(5), 499–500. http://doi.org/10.1038/nn1223

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17. doi:10.1016/0010-0285(70)90002-2

Shepard, R. N., & Cooper, L. A. (1992). Representations of color in the blind, color-blind, and normally sighted. *Psychological Science*, *3*, 97–104.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751–754.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–89. doi:10.1006/nimg.2001.0978

Unger, L., Fisher, A. V., Ventura, S., Nugent, R., & MacLellan, C. J. (2016). Developmental changes in the semantic organization of living kinds. *Journal of Experimental Child Psychology*, *146*, 202–222.

Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, *10*(11), 792–802. doi: 10.1038/nrn2733

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, *78*, 569–576. doi:10.1080/01621459.1983.10478009

Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. Human Brain Mapping, 31, 1459–1468. doi:10.1002/hbm.20950

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia*, *47*(3), 859–868. doi:10.1016/j.neuropsychologia.2008.12.029

Wang, J., Baucom, L. B., & Shinkareva, S. V. (2013). Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*, *34*(5), 1133–1147. doi:10.1002/hbm.21498

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, Teacher, and Stethoscope: Neural Representation of Different Types of Semantic Relations. *The Journal of Neuroscience*, *38*(13), 3303-3317. doi: 0.1523/JNEUROSCI.2562-17.2018

Yee, E., & Thompson-Schill, S. L. (2016). Putting Concepts into Context. *Psychonomic Bulletin & Review*, 1–13. doi:10.3758/s13423-015-0948-7