Similarity and Induction[*]

Matthew Weber
Princeton University

Daniel Osherson
Princeton University

October 1, 2008

Abstract

We advance a theory of inductive reasoning based on similarity, and test it on arguments involving mammal categories. To measure similarity, we quantified the overlap of neural activation in left Brodmann area 19 and the left ventral temporal cortex in response to pictures of different categories; the choice of of these regions is motivated by previous literature. The theory was tested against probability judgments for 40 arguments generated from 9 mammal categories and a common predicate. The results are interpreted in the context of Hume's thesis relating similarity to inductive inference.

Introduction

David Hume (1748) famously asserted a role for similarity in non-deductive inference. Here is the well-known passage.

[*] We thank Sergey Blok, James Haxby, Douglas Medin, and Lawrence Parsons for discussion and assistance in various stages of this work. Alvin Goldman and Adam Elga, and Scott Weinstein offered comments on earlier drafts. Two anonymous referees graciously provided helpful suggestions. Osherson was supported by a grant from the Henry Luce Foundation. Contact: matthew/osherson@princeton.edu.

> “In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects. ... From causes which appear similar we expect similar effects.”

Hume’s view is consistent with his predecessor Locke (1689), for whom analogy was “the great rule of probability.” Just what Locke and Hume meant by the term “probability” is open to discussion, but their thesis is clear. *Similarity often lies behind inductive inference.* The goal of the present essay is to sharpen this insight.

By “induction” we’ll understand a certain relation between a list of statements and some further statement. The first statements are called *premises*, the last the *conclusion*, and the ensemble an *argument*. The *inductive strength* of an argument for a given person will be identified with the subjective conditional probability she attaches to the conclusion given the premises. This definition raises questions about subjective probability in the minds of people who misunderstand chance. (Most college students can be led to incoherent estimates of probability; see Bonini et al., 2004; Tentori et al., 2004.) So we will just assume that the probability idiom conveys a familiar kind of *psychological coherence condition*. An argument is strong to the extent that the reasoner would *find it odd* to believe the premises without believing the conclusion. Squeezing this mental sensation into the unit interval and calling it probability provides a rough measure.

Inductive inferences are of such diverse character that we may despair of treating them within a unified theory. It will make things easier to limit attention to premises and conclusion with subject-predicate syntax, the same predicate appearing throughout. Here is an illustration along with the style of abbreviation used in the sequel.

- *Premise*: Sheep have at least 18% of their cortex in the frontal lobe.
- *Conclusion*: Goats have at least 18% of their cortex in the frontal lobe.

Abbreviation:

- *Premise*: Qa.
- *Conclusion*: Qc.

a and c are sheep and goats, respectively. Q is the common predicate.

In the general case there may be multiple premises.

To predict the conditional probability of the conclusion given the premises, it is necessary to start from some sort of information. We assume that two kinds of quantity are available. First, we give ourselves the *similarity* between all pairs of objects in play. Second, we give ourselves all relevant *unconditional* probabilities. Similarity will be assumed to take values in the unit interval, be symmetric, and return 1 in case of identity. Denoting the similarity function by $\mathit{similarity}(\cdot, \cdot)$, our assumptions are thus:

Given objects a, b:

- $\mathit{similarity}(a, b) \in [0, 1]$
- $\mathit{similarity}(a, b) = \mathit{similarity}(b, a)$
- $\mathit{similarity}(a, a) = 1$.

As mentioned, we also help ourselves to the *unconditional* probabilities of each premise and the conclusion of an argument. All that's missing is the *conditional* probability of the conclusion given the premises, in other words, the inductive strength of the argument. Our project is thus to forge conditional probability from unconditional probability plus similarity. The criterion of success will be conformity to the estimates of conditional probability that people typically offer. This puts a descriptive spin on Hume's thesis, which is consistent with his doubts about the normative justification of induction.

An algorithm for constructing conditional probability

Bayes' Theorem has been useful in modeling many aspects of inductive judgment (e.g., Heit, 1998; Tenenbaum et al., 2007) but offers little help in the present enterprise. For it expands the desired conditional probability into an expression featuring a *different* conditional probability, just as challenging (namely, the likelihood):

Bayes' Theorem:

For an argument Qa / Qc with a single premise Qa:

$$\textit{Prob}\,(Qc \mid \; Qa) = \frac{\textit{Prob}\,(Qa \mid \; Qc) \times \textit{Prob}\,(Qc)}{\textit{Prob}\,(Qa)}$$

We get more help from the usual *definition* of conditional probability since it suggests that we attempt to estimate the probability of conjunctions of statements.

Usual definition of conditional probability:

For an argument $Qa_1 \cdots Qa_n$ / Qc:

$$\textit{Prob}\,(Qc \mid \; Qa_1 \cdots Qa_n) = \frac{\textit{Prob}\,(Qc \,\&\, Qa_1 \,\&\, \cdots \,\&\, Qa_n)}{\textit{Prob}\,(Qa_1 \,\&\, \cdots \,\&\, Qa_n)}$$

Following the conjunction strategy, our problem reduces to this one: We're given the probabilities of the conjuncts, for example, the probability that sheep have at least 18% of their cortex in the frontal lobe. We're also told the pairwise similarity among $\{a_1 \cdots a_n, c\}$ for example, the similarity of goats and sheep. From this we wish to construct a sensible estimate of the probability of the conjunction. This is all we need to derive conditional probability.

So what is a sensible estimate of the probability of a conjunction $Qb_1 \,\&\, \cdots \,\&\, Qb_n$?[1] One constraint is that the conjunction probability fall between the minimum and maximum allowed by the laws of chance. It can be shown (Neapolitan, 1990) that the lowest possible

[1] The model presented here is an alternative to "QPf," described in Blok et al. (2007a). It relies on some of the same concepts. Other attempts to relate similarity and conditional probability include Blok et al. (2007b).

value is $\max\{0,\ 1-n+\sum_{i=1}^{n} \textit{Prob}\,(Qb_i)\}$, that is, one plus the sum of the n conjunct probabilities minus n, or zero if the latter number is negative. The highest possible value is the minimum probability of the conjuncts. Thus, we have:

$$\boxed{\begin{aligned}&\max\{0,\ 1-n+\sum_{i=1}^{n} \textit{Prob}\,(Qb_i)\} \quad \leq \\ &\textit{Prob}\,(Qb_1 \;\&\; \cdots \;\&\; Qb_n) \quad \leq \\ &\min\{\textit{Prob}\,(Qb_1), \cdots, \textit{Prob}\,(Qb_n)\}\end{aligned}}$$

For example, if $\textit{Prob}\,(Qa) = .8$ and $\textit{Prob}\,(Qb) = .4$ then the probability of the conjunction cannot exceed .4, and it can't fall below .2:

$$.2 \;=\; 1-2+.8+.4 \;\leq\; \textit{Prob}\,(Qa \;\&\; Qb) \;\leq\; \min\{.4, .8\} \;=\; .4.$$

Now notice that if a and b were identical — had maximal similarity — then Qa and Qb would express the same proposition. So the conjunction of Qa with Qb would be logically equivalent to Qa, and also to Qb. In this case, the probability of the conjunction would fall at the upper bound of the possible interval, namely, the minimum of the conjunct-probabilities. In contrast, the conjunction makes a stronger claim to the extent that a and b are *dis*similar, so low similarity should situate the probability of the conjunction closer to the bottom of the permitted interval. We allow the similarity of a and b to determine a position between these extremes. The conjunction seen earlier serves as illustration. Recall that its probability must fall in the interval from .2 to .4. The probability we construct for it is the weighted sum of these endpoints, where the weights are given by the similarity of a to b. Summarizing this example:

Probability constructed for $Qa \;\&\; Qb$: Suppose that $\textit{Prob}\,(Qa) = .8$ and $\textit{Prob}\,(Qb) = .4$ so that the smallest possible value of $\textit{Prob}\,(Qa \;\&\; Qb)$ is .2, and the greatest possible value of $\textit{Prob}\,(Qa \;\&\; Qb)$ is .4. Then the value attributed to $\textit{Prob}\,(Qa \;\&\; Qb)$ is:

$$[.2 \times (1 - \textit{similarity}(a,b))] + [.4 \times \textit{similarity}(a,b)].$$

In the general case we're given a conjunction with n conjuncts. As usual, we assume that we know the unconditional probabilities of its conjuncts. This allows us to compute the lower and upper bounds on its possible probability. We take the similarity-factor to be the minimum of all the pairwise similarities among the objects appearing in the conjunction. Then we estimate the probability of the conjunction to be the weighted sum of the lower and upper bounds, where similarity controls the weights. That is:

Constructed probability (general case): We're given the conjunction $\mathbf{C} = Qb_1 \ \& \ \cdots \ \& \ Qb_n$.

- Let $\mathbf{p} = \max\{0,\ 1 - n + \sum_{i=1}^{n} \textit{Prob}\,(Qb_i)\}$, the least possible value of **C**.
- Let $\mathbf{P} = \min\{\textit{Prob}\,(Qb_1), \cdots, \textit{Prob}\,(Qb_n)\}$, the greatest possible value of **C**.
- Let $\mathbf{sim} = \min\{\textit{similarity}(b_i, b_j) \mid \ i, j \leq n\}$.

Then we set:

$$\textit{Prob}\,(\mathbf{C}) = [\mathbf{p} \times (1 - \mathbf{sim})] + [\mathbf{P} \times \mathbf{sim}].$$

As a sanity check, it can be verified that our scheme satisfies the conjunction law:

$$\textit{Prob}\,(Qb_1 \ \& \ \cdots \ \& \ Qb_n) \ \geq \ \textit{Prob}\,(Qb_1 \ \& \ \cdots \ \& \ Qb_n \ \& \ Qb_{n+1})$$

This law helps to motivates the use of minimum similarity in our model; for, it would not be satisfied had we relied on average or maximum instead.

We must also consider conditional probabilities that involve negated statements. They arise from negated premises or conclusion, as illustrated here:

$$\textit{Prob}\,(Qc \mid \ \neg Qa) = \frac{\textit{Prob}\,(Qc \ \& \ \neg Qa)}{\textit{Prob}\,(\neg Qa)}$$

In this case, we use the same procedure as before except that we substitute one minus the unconditional probability for negated statements. We also use one minus the similarity of two objects that appear in statements of opposite polarity. For example, the high similarity of lions to cougars should lower the probability of $Q(\text{lion})\ \&\ \neg Q(\text{cougar})$ and raise the probability of $\neg Q(\text{lion})\ \&\ \neg Q(\text{cougar})$.

With these principles, probabilities may be constructed for arbitrary conjunctions. Lower and upper bounds are computed as before, after negations are transformed by one-minus. The minimum pairwise similarity is also determined, using the one-minus operation on similarities associated with conjuncts of opposite polarity. Then a position in the permitted interval is selected via a convex sum, and this is taken to be the probability of the conjunction. To illustrate:

Constructing probability for the conjunction:

$$Qb_1\ \&\ \neg Qb_2\ \&\ \neg Qb_3\ \&\ Qb_4$$

First compute lower and upper bounds (**p**, **P**) for the probability of the conjunction as before but after substituting $1 - \textit{Prob}\,(Qb)$ for $\textit{Prob}\,(\neg Qb)$. Next compute the minimize pairwise similarity (**sim**) among $\{b_1, b_2, b_3, b_4\}$, using the one-minus operation on similarities associated with conjuncts of opposite polarity. Then the probability of the conjunction is taken to be:

$$[\mathbf{p} \times (1 - \mathbf{sim})] + [\mathbf{P} \times \mathbf{sim}].$$

Our method satisfies important coherence conditions involving upper and lower bounds along with the conjunction law, as seen above. Moreover, it is easy to show that the method assigns zero probability to contradictions. That is, according to the scheme described above:

$$\textit{Prob}\,(Qb_1\ \&\ \ldots\ \&\ Qa\ \&\ \ldots\ \&\ \neg Qa\ \&\ \ldots\ Qb_n) = 0.$$

Satisfaction of this principle once again relies on the role of minimum similarity in our model; use of average or maximum would assign positive probability to some contradictions.

The model also exhibits the *premise diversity effect* inasmuch as $\mathit{Prob}\,(Qc \mid \; Qb_1, Qb_2)$ tends to be greater for smaller values of $\mathit{similarity}(b_1, b_2)$.[2] Diversity of premises is often claimed to entail greater inductive strength (e.g., Hempel, 1966; Franklin and Howson, 1984). The psychological counterpart of this thesis is explored in Osherson et al. (1990); López et al. (1997); López (1995); Choi et al. (1997), and elsewhere (for a review, see Heit et al., 2005).

On the other hand, there is no guarantee that the conjunctions over a given set of statements are assigned probabilities that sum to one. For example, here are all eight conjunctions over the three statements Qc, Qa, Qb.

$Qc \;\&\; Qa \;\&\; Qb$	$Qc \;\&\; Qa \;\&\; \neg Qb$
$Qc \;\&\; \neg Qa \;\&\; Qb$	$\neg Qc \;\&\; Qa \;\&\; Qb$
$\neg Qc \;\&\; \neg Qa \;\&\; Qb$	$\neg Qc \;\&\; Qa \;\&\; \neg Qb$
$Qc \;\&\; \neg Qa \;\&\; \neg Qb$	$\neg Qc \;\&\; \neg Qa \;\&\; \neg Qb$

Our method does not in general assign them eight numbers that sum to unity. The matter can be rectified through normalization but there is no need for this in the present context because conditional probabilities are ratios and normalizing has no numerical impact. Thus, to construct $\mathit{Prob}\,(Qc \mid \; Qa, \neg Qb)$, we calculate the probabilities of $Qc \;\&\; Qa \;\&\; \neg Qb$ and $\neg Qc \;\&\; Qa \;\&\; \neg Qb$, then insert them into the usual formula for conditional probability (equivalent to the definition shown earlier):

$$\mathit{Prob}\,(Qc \mid \; Qa, \neg Qb) = \frac{\mathit{Prob}\,(Qc \;\&\; Qa \;\&\; \neg Qb)}{\mathit{Prob}\,(Qc \;\&\; Qa \;\&\; \neg Qb) + \mathit{Prob}\,(\neg Qc \;\&\; Qa \;\&\; \neg Qb)}.$$

It remains to test whether the scheme just presented approximates human intuition about chance.

[2] This holds only if the similarity of b_1 to c, and of b_2 to c are not too high. When the latter two similarities approach unity, so should the similarity between b_1 and b_2 (in the limit, similarity becomes identity, which is transitive). Small values of $\mathit{similarity}(b_1, b_2)$ are thus difficult to interpret if both $\mathit{similarity}(b_1, c)$ and $\mathit{similarity}(b_2, c)$ are high.

Behavioral data

Eliciting estimates of probability

To test our theory, twenty undergraduates were asked to reason about these categories

Bears	Camels	Cougars
Dolphins	Elephants	Giraffes
Hippos	Horses	Lions

and the predicate (Q): *have at least 18% of their cortex in the frontal lobe*. The students offered probabilities for 40 arguments with the forms shown here.

Form	Number of instances
$\mathit{Prob}\,(Qc \mid\ Qa)$	5
$\mathit{Prob}\,(Qc \mid\ \neg Qa)$	5
$\mathit{Prob}\,(\neg Qc \mid\ Qa)$	5
$\mathit{Prob}\,(\neg Qc \mid\ \neg Qa)$	5
$\mathit{Prob}\,(Qc \mid\ Qa, Qb)$	10
$\mathit{Prob}\,(Qc \mid\ Qa, \neg Qb)$	10

They also estimated the nine unconditional probabilities corresponding to the nine mammal categories.[3] For example, they were asked:

Sample requests for probability estimates:

the probability that cougars have at least 18% of their cortex in the frontal lobe given that this is the case for lions.

the probability that horses have at least 18% of their cortex in the frontal lobe given that this is the case for elephants but not for giraffes.

the probability that bears have at least 18% of their cortex in the frontal lobe.

[3] Participants were interviewed singly. Questions were posed in individualized random order via computer interface. Responses were made with a slider that controlled a field displaying numbers in the unit interval. The concept of conditional probability was reviewed prior to testing.

The probabilities obtained were averaged across participants, yielding the numbers displayed in Table 1.

Then we attempted to predict the conditional probabilities on the basis of similarity plus the unconditional probabilities, using the scheme described above. The unconditional probabilities are available from the data, having been directly elicited. But what shall we use as our measure of similarity?

Similarity untainted by inductive inference

We could ask the students to provide numerical estimates of the similarity of pairs of species, using a rating scale. But such a procedure would not fairly test Hume's idea. His thesis was that perceived similarity gives rise to judged probability. We must not inadvertently test the converse idea, that perceived probability gives rise to judged similarity. After all, it could be that lions and cougars seem similar because inferences from one to the other strike us as plausible.[4] Then similarity would indeed be related to induction but not in the way Hume intended. Basing similarity on the set of features (e.g., "nocturnal") that students identify in different species raises the same risk of circularity inasmuch as such judgments may involve probability (e.g., the probability that cougars are nocturnal assuming that lions are, or assuming that cougars have large frontal lobes).[5] To focus on Hume's idea, we need to operationalize similarity without allowing probability estimates to play an implicit role.

For this purpose, we adopt the idea that similarity of categories — like *horses* and *camels* — is determined by their respective neural representations. To quantify neural similarity, we rely on functional magnetic resonance imaging (fMRI) to identify the patterns of activation that support the categories; proximity of categories is then measured in physical terms.

[4] The influence on similarity judgments of confirmed theory is discussed in Earman (1992); Wayne (1995).

[5] Krantz et al. (1989); Smith et al. (1993); Sloman (1993) and Tenenbaum et al. (2007) offer alternative perspectives on the use of features to explain similarity and inference.

Neurophysiological data

Obtaining activation maps

Twelve new subjects were recruited to perform a classification task during scanning. On each trial, they viewed a category label for one of the nine mammals used in the behavioral study on probability estimates. Then they viewed a series of images of the designated mammal. But occasionally there was an intruder that had to be signaled via button-press. See Figure 1. These “catch trials” were intended to ensure that the images were processed; only trials without intruders were analyzed further. There was also a series of control trials substituting phase-scrambled versions of the original images (hence, unidentifiable but with the same spatial frequencies and overall luminance). In the latter trials, subjects searched for a low-contrast cross hatch (#), present only in catch trials (excluded from analysis). Details of the procedure and analysis are provided in the Appendix.

None of the fMRI subjects participated in the probability assessments. Also, no mention was made of similarity or probability either before or during scanning. The fMRI subjects simply verified the category of mammal images (or verified in control trials that # was absent).

The fMRI procedure parcels the brain into roughly $50,000$ cubes called *voxels*, 3 millimeters on a side. For each voxel, we obtained a measure of the metabolic activity provoked by recognizing bears, another value for giraffes, and so forth. The measure is the β coefficient for a given mammal’s regressor in the best linear model of the voxel’s behavior in the experiment; see the Appendix. These values were averaged across the 12 subjects (after projection of each brain onto a common template). Average activations were also obtained when viewing phase-scrambled pictures of each mammal. For each mammal, the activations arising from viewing its scrambled version were subtracted from the activations produced by the verification task. The resulting distribution of corrected values (obtained from the subtraction) induces a “map” of activations over the brain. There is one such map for each mammal. We compared the maps for each pair of mammals to estimate similarity. The method of comparison will be explained shortly.

Choosing neural regions

First we address the question: which structure of the brain should be mapped, that is, where are mammal categories located? It has been observed that lesions to the left temporal lobe are sometimes associated with specific deficits in knowledge of biological categories including mammals, vegetables, and fruit, sparing knowledge of human artifacts like furniture and tools (Warrington and Shallice, 1984; Saffran and Schwartz, 1994; Capitani et al., 2003). Also, single cell recording from inferior temporal cortex in monkeys reveals neurons that are responsive to natural categories (although their specificity may be influenced by size and position in the visual field, among other features; see Zoccolan et al., 2007). Partially converging information is available from human neuroimaging. A review of studies by Martin (2001) points to activity (often bilateral) in the lateral fusiform gyrus, medial occipital cortex, and superior temporal sulcus when subjects are asked to identify and name pictures of animals. Inferior regions of the left occipital cortex seem also to be recruited when viewing pictures of animals in contrast to tools (Martin et al., 1996). Broadly consistent findings emerge from studies of processing category-names (e.g., Perani et al., 1999, who report left fusiform gyrus activations for animals). Kounios et al. (2003) reach similar conclusions in their summary of the literature. There are, however, many inconsistent findings in both the clinical and neuroimaging literature (Caramazza, 2000; Joseph, 2001; Gerlach, 2007). It is also unclear whether any given brain locus holds an integrated animal representation rather than perceptual or abstract features associated with it (e.g., visual properties in the left fusiform gyrus; see Thompson-Schill et al., 2006).

Moreover, structures beyond the temporal and occipital lobes have been implicated in the manipulation of conceptual knowledge. For example, Freedman et al. (2001) report categorical responding to pictures of cats and dogs by neurons in the lateral prefrontal cortex of monkeys.[6] Human neuropsychology and neuroimaging also implicate the premotor cortex of the frontal lobe in object categorization, especially of manipulable objects such as fruit,

[6] The same categoricity, however, was observed when the monkeys were trained on concepts involving cat/dog mixtures. Note that the LPFC is directly interconnected with inferior temporal cortex (Webster et al., 1994).

tools, and clothing, although findings are not entirely consistent (e.g. Martin et al., 1996; Chao et al., 1999; Gerlach et al., 2002; for reviews, see Gainotti, 2000 and Martin, 2007). There has been no similar report for mammal categories.

Consonant with the appearance in the previous literature of both temporal and occipital structures underlying the representation of mammals, we investigated two broad regions of the brain that include some of the principal areas discussed above. One region is left hemispheric Brodmann Area 19 (left BA19), overlapping the lateral occipital gyri. The other is left ventral temporal cortex (left VTC) comprising the inferior temporal gyrus, fusiform gyrus, and the parahippocampal gyrus. See Figure 2.

Comparing activation maps

Suppose that R is one of our two regions, comprised of voxels $v_1 \cdots v_n$. Each v_i has a level of activation h_i for *horse*, an activation c_i for *camel*, and so forth. Then $\Sigma(h_i - c_i)^2$ (the sum of squared deviations) is a natural measure of the dissimilarity in R of the respective neural representations of horses and camels. The calculation of this sum can be summarized as follows.

Voxel	Activation for *horse*	Activation for *camel*	Squared deviation
v_1	h_1	c_1	$(h_1 - c_1)^2$
v_2	h_2	c_2	$(h_2 - c_2)^2$
v_3	h_3	c_3	$(h_3 - c_3)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
			Sum of squared deviations

The nine mammals give rise to $36 = \binom{9}{2}$ such computations of dissimilarity. To convert them to similarity, each is first inverted (divided into 1). Then the 36 resulting numbers are linearly scaled to run from $\frac{1}{3}$ to $\frac{2}{3}$. Occupying just the middle of the unit interval leaves room for pairs less similar than ours (e.g., moles compared to dolphins), as well as pairs more similar (e.g., camels versus dromedaries).

Each of the two neural regions yields its own set of 36 similarities via the foregoing procedure. They are shown in Table 2. Corresponding similarities are highly correlated ($r = .840$, $N = 36$), suggesting a common mental reality.[7]

Predicting conditional probabilities

Our neural measure is uncontaminated by use of strength-of-inference as an index of similarity; for, the neural measure was obtained from mammal-stimuli individually, with no mention of similarity or probability. Relative to the model of inductive strength advanced above, a pure test of Hume's thesis is therefore possible. It suffices to enter neural similarity into the model, along with the unconditional probabilities culled directly from subjects. The predictions generated thereby can then be compared to the results of the probability experiment.

Specifically, the experiment produced 40 numbers, corresponding to the arguments shown in Table 1. Each is an average estimate of conditional probability, to be paired with the corresponding probability calculated from our model based on neural similarity. Because similarity was calculated twice (on the basis of two neural regions), predictions are evaluated separately for left VTC and left BA19.

The relation between predictions and observations is plotted in Figure 3 for each region. A reliable linear association is discernable in each case (Pearson $r = .716, .728$, for left VTC and left BA19, respectively).

The predictive accuracy of the model cannot be attributed exclusively to the use of neural similarity; the model also rests on estimates of unconditional probability, and these were

[7] The results reported below are virtually identical if the activations for each mammal in a given region are "mean-centered." To mean-center mammal M in region R, the average activation (β) in the map for M in R is subtracted from all the activations in the map prior to computing the sum of squared deviations between M and any other mammal.

obtained from the same subjects whose conditional probabilities are at issue. To isolate the role of similarity, we substituted for neural similarity 36 random numbers drawn uniformly from $\left[\frac{1}{3}, \frac{2}{3}\right]$. In 1000 random trials, the average correlation between predicted and observed conditional probabilities was $.402$ (SD = $.143$). Only one random trial reached $r = .716$.

Subsequent to fMRI scanning, subjects rated the similarity of all 36 pairs of mammals on a scale from 0 to 100%. The ratings for each pair were then averaged across the 12 subjects. When rated similarity is inserted into the model, the correlation between predicted and observed estimates of probability is $.675$, slightly lower than the correlations based on left VTC and left BA19. The correlation between rated similarity and similarity based on left VTC is $.655$. The correlation is $.648$ between rated similarity and similarity based on left BA19.

Discussion

Other avenues to neural similarity

We calculated similarity from left VTC and left BA 19 because these areas are suggested by the previous literature devoted to the neural representation of natural categories like animals. Certain other areas, however, yield equally good results. For example, when similarities are computed from the left primary visual cortex, the correlation between predicted and observed estimates of conditional probability is $.707$.

Squared deviation is perhaps the simplest approach to neural similarity but at least one other technique works as well. Given a neural region R and mammal m, the alternative assigns a point in three-dimensional space that reflects the overall position of the activations in R in response to m. The similarity between two mammals is then measured by the Euclidean distance between the points assigned to them (with inversion and linearization to $\left[\frac{1}{3}, \frac{2}{3}\right]$, as before). Used in the rhinal sulcus of the temporal lobe, this index of similarity predicts conditional probability at $r = .728$. On the other hand, most other neural structures have no predictive success under either of the approaches to similarity discussed here. Understanding how and where similarity is coded in the brain is a topic of current investigation

(Weber et al., in preparation).

A lower bound for conjunctions based on independence

In our theory of inductive strength, the value of $\mathit{Prob}\,(Qb_1 \,\&\, \cdots \,\&\, Qb_n)$ is situated in the interval from $\max\{0,\ 1-n+\sum_{i=1}^{n} \mathit{Prob}\,(Qb_i)\}$ to $\min\{\mathit{Prob}\,(Qb_1), \cdots, \mathit{Prob}\,(Qb_n)\}$. Similarity is used to choose a point in the interval, with low similarity pushing the point to the lower bound. Let us consider changing the lower bound to the product of the probabilities of the conjuncts: $\prod_{i=1}^{n} \mathit{Prob}\,(Qb_i)$. The latter bound embodies the idea that low similarity signals the stochastic independence of the conjuncts rather than their incompatibility. This might correspond better to what Hume had in mind since he seems to take the absence of similarity to reflect no reason for belief rather than reason for disbelief (Cohen, 1980).

It is therefore worth reporting that the revised model with multiplicative lower bound underperforms the original model. In every neural region examined (and with both measures of similarity), correlations between observed and predicted probabilities are about 0.1 lower for the revised model. Also, the independence bound implies $\mathit{Prob}\,(Qb \,\&\, \neg Qb) > 0$ whenever $0 < \mathit{Prob}\,(Qb) < 1$, a coherence violation.

Limitations and extensions

Hume's thesis about induction has here been examined through the lens of a particular model of probability judgment, which starts from unconditional probability and pairwise similarity. The model cannot generate arbitrary distributions. A joint distribution over the statements $Qb_1, Qb_2 \cdots\ Qb_n$ (where Q is a predicate and $b_1 \cdots\ b_n$ are objects) requires $2^n - 1$ numbers in general. The scheme described here specifies distributions based on only $n + \binom{n}{2}$ numbers (n unconditional probabilities and all pairwise similarities). It follows that our method must omit many potential distributions. This kind of compression, however, may be compatible with describing aspects of *human judgment*, which likely chooses distributions from a limited set in most circumstances.[8]

Even if our method yields distributions that describe human judgment, we have not provided evidence that reasoning proceeds by constructing probabilities for conjunctions. Without such evidence, our model should be interpreted as describing just an input-output relation (unconditional probabilities and similarities in, conditional probability out).

Further progress in constructing a psychological theory will require distinctions among predicates. The predicates in the following list are adapted to the present model because they have biological content without evoking detailed knowledge in the minds of most college students.

- have at least 18% of their cortex in the frontal lobe
- have trichromatic vision
- can suffer muscle damage through contact with poliomyelitis
- brain/body mass ratio is 2 percent or more
- require at least 5 hours of sleep per day for normal functioning
- sex drive varies seasonally
- have muscle-to-fat ratio at least 10-to-1
- spends at least half its time foraging for food
- have testosterone levels at least 10 times higher in males compared to females

They are also "shareable," in contrast to the predicate *eats more than half the grass in its habitat*, which can't be true of different species in the same habitat. Our model needs adjustment in the face of non-shareability.

We must also acknowledge the difficult question of how the brain generates different similarity functions relevant to distinct predicates. It is often noted that lynx and house cats, for example, are perceived as similar in the context of biological predicates (like *have trichromatic vision*) but dissimilar in the context of economic variables (like *are sold in*

[8] A natural generalization of our model replaces (binary) similarity with the *homogeneity* of sets of $k \geq 2$ objects. To illustrate, such a measure might assign greater homogeneity to { camels, horses, giraffes } compared to { camels, bears, lions }. All distributions over $Qb_1, Qb_2 \cdots Qb_n$ can be generated by a model like ours that relies on n-ary homogeneity.

most pet shops; see Osherson et al., 1986, and Medin et al., 1993, for discussion). Perhaps biological similarity is fundamental to our conception of natural categories like mammals, with other kinds of similarity (e.g., economic) resulting from its interaction with beliefs about specific mechanisms.

Invoking “mechanism” brings to mind distinctions about the causes Hume might have been referring to in the passage quoted above. We here interpret species as causing their properties, but narrower conceptions (e.g., based on the physical interaction of parts) might be associated with a different psychology. It is known that much of ordinary reasoning relies on causal schemas in a general sense (Rehder, 2006, 2007; Sloman, 2005). It is thus noteworthy that our model is easily adapted to represent constraints on the distribution of properties (constraints that might arise from causal knowledge). This is achieved by setting the probability of specific conjunctions to zero, or imposing independence relations among them.

Other challenges arise when arguments display distinct predicates in premise and conclusion, or involve relations like *preys on*. Inferences involving non-natural kinds — like artifacts and political parties — bring fresh distinctions to light. Accommodation is also needed for the tendency of even well-educated respondents to issue probabilistically incoherent estimates of chance, or to judge similarity asymmetrically (Tversky, 1977; but see also Aguilar and Medin, 1999). Confronting these complexities is inevitable for the development of any theory of human inductive judgment. The data presented here suggest merely that progress will involve similarity in something like the sense Hume had in mind.

Hume

In the foregoing discussion we have interpreted Hume’s thesis as a psychological claim, namely, that inductive inference (as people actually perform it) is driven by similarity. Our formal rendition of this claim enriches the determinants of inductive strength by appealing to the prior probabilities of premises and conclusion. Such additions notwithstanding, the predictive success of the model (limited though it be) supports Hume’s thesis.

To test the thesis as Hume intended it, we relied on a measure of similarity that is free from contamination by inferential reasoning. The measure rests on comparison of the neural representations of mammal-categories, in the absence of judgments of similarity or probability. This is not to deny that over many years, inferences about the properties of mammals might affect how they are ultimately coded in the brain. Thus, lions and cougars may be represented be common patterns because they are perceived to share many properties. Neural similarity could therefore depend on inference via this route. Nonetheless, our measure of similarity is directly mediated by the mental representation of concepts rather than accessing the machinery of inductive cognition. This seems a fair way of making Hume's claim precise. So perhaps our results sustain his claim that *similarity provides a partial foundation for understanding inductive inference*. Whether similarity could *justify* an inductive inference remains a separate matter.

Appendix: fMRI details

Image acquisition

Scanning was performed with a 3-Tesla Siemens Allegra fMRI scanner. Participants' anatomical data were acquired with an MPRAGE pulse sequence (160 sagittal slices) before functional scanning. Functional images were acquired using a T2-weighted echo-planar pulse sequence with 33 64×64-voxel slices, rotated back by 5 degrees on the left-right axis (axial-coronal $-5°$). Voxel size was $3 \times 3 \times 3$ mm, with a 1-mm gap between slices. The phase encoding direction was anterior-to-posterior. TR was 2000 ms; time to echo was 30 ms; flip angle was $90°$. Field of view was 192×192 cm.

fMRI task

During scanning, participants performed several trials of an *experimental task* on intact stimuli and a *visual baseline* task on phase-scrambled stimuli. During a trial of the experimental task, participants saw the name of one of our nine mammals (bear, camel, cougar, dolphin, elephant, giraffe, hippo, horse, lion) for 2 s, then a series of 8 intact mammal images presented for 1.5 s each (totaling 12 s of images), then a question mark. Participants were instructed to press a key at the question mark just in case any of the presented images did not match the presented name. Thus, if the word "bear" was followed by 8 bear images, a key-press was not required at the end of the trial; in contrast, a key-press was expected if (e.g.) a zebra intruded in the sequence of bears. The form of the visual baseline task was identical, except that no word was presented before the images, and participants searched for a low-contrast crosshatch (#) in the sequence instead of a category mismatch. The images in each baseline trial were phase-scrambled versions of one of the mammals; thus there was a baseline trial consisting only of scrambled bears, one consisting only of scrambled camels, etc.

The study was organized into 7 runs of the experimental task, with 9 trials each (one for each mammal). There was also 1 run of the baseline task, with 11 trials (one trial without #

for each mammal plus two with #). Each run contained 0, 1, or 2 trials requiring a response; these trials were discarded, and only trials in which participants verified an unbroken stream of intact or scrambled images were analyzed. (Discarded trials thus served to verify the subject's attention; performance on them was in fact perfect.)

Image analysis

Functional data were registered to the participant's anatomical MRI, despiked, smoothed with a 6-mm full-width at half-max Gaussian kernel, and normalized to percent signal change. For each participant, multiple regression was used to generate β values representing each voxel's activity in each mammal condition and each visual baseline condition. To calculate the β's, all variables were convolved with a canonical, double gamma hemodynamic response function and entered into a general linear model. Motion estimates were included as regressors of no interest; trials requiring a key-press were discarded. In a given voxel, the activation level for mammal m was then defined as the β for m's mammal condition minus the β for m's visual baseline. We relied on the statistical package AFNI (Cox, 1996) for preprocessing.

The 12 resulting β maps for a given mammal (one for each subject) were projected into Talairach space and averaged, leaving us with nine such maps, one for each mammal. Only voxels present in the intersection of all participants' intracranial masks were considered. These average activation maps were the input to all subsequent analyses. Brodmann areas were identified by application of the *MRIcro atlas* (Rorden and Brett, 2000).

References

C. M. Aguilar and D. L. Medin. Asymmetries of comparison. *Psychonomic Bulletin & Review*, 6: 328–337, 1999.

S. Blok, D. Medin, and D. Osherson. From similarity to chance. In Evan Heit and Aidan Feeney, editors, *Inductive Reasoning: Experimental, Developmental, and Computational Approaches*. Cambridge University Press, 2007a.

Sergey Blok, Douglas Medin, and Daniel Osherson. Induction as conditional probability judgment. *Memory & Cognition*, 35(6):1353–1364, 2007b.

N. Bonini, K. Tentori, and D. Osherson. A different conjunction fallacy. *Mind and Language*, 19 (2):199 210, 2004.

E. Capitani, M. Laiacona, B. Mahon, and A. Caramazza. What are the facts of semantic category-specific deficits? a critical review of the evidence. *Cognitive Neuropsychology*, 20:213–261, 2003.

A. Caramazza. The organization of conceptual knowledge in the brain. In M. S. Gazzaniga, editor, *The New Cognitive Neurosciences*, pages 1037–1046. MIT Press, Cambridge MA, 2000.

L. L. Chao, J. V. Haxby, and A. Martin. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, 2:913–919, 1999.

I. Choi, R. E. Nisbett, and E. E. Smith. Culture, Categorization and Inductive Reasoning. *Cognition*, 65:15 – 32, 1997.

L. J. Cohen. Some Historical Remarks on the Baconian Conception of Probability. *Journal of the History of Ideas*, 41(2):219 – 231, 1980.

R. W. Cox. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomed. Res.*, 29:162–73, 1996.

J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge MA, 1992.

A. Franklin and C. Howson. Why Do Scientists Prefer to Vary Their Experiments? *Studies in the History and Philosophy of Science*, 15:51–62, 1984.

D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–6, 2001.

G. Gainotti. What the locus of brain lesion tells us about the nature of the cognitive defect underlying category-specific disorders: A review. *Cortex*, 36:539–559, 2000.

C. Gerlach. A review of functional imaging studies on category specificity. *Journal of Cognitive Neuroscience*, 19(2):296–314, 2007.

C. Gerlach, I. Law, and O. B. Paulson. When action turns into words. Activation of motor-based knowledge during categorization of manipulable objects. *Journal of Cognitive Neuroscience*, 14 (8):1230–1239, 2002.

E. Heit. A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford and N. Chater, editors, *Rational Models of Cognition*, pages 248 – 274. Oxford University Press, Oxford UK, 1998.

E. Heit, U. Hahn, and A. Feeney. Defending diversity. In W. Ahn, R. Goldstone, B. Love, A. Markman, and P. Wolff, editors, *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin*, pages 87–99. American Psychological Association, Washington DC, 2005.

C. G. Hempel. *Philosophy of Natural Science*. Prentice Hall, Englewood Cliffs NJ, 1966.

David Hume. *An Enquiry Concerning Human Understanding*. (unknown to me), 1748.

J .E. Joseph. Functional neuroimaging studies of category specificity in object recognition: A critical review and meta-analysis. *Cognit., Affect. Behav. Neurosci.*, 1:119–136, 2001.

J. Kounios, P. Koenig, G. Glosser, C. DeVita, K. Dennis, P. Moore, and M. Grossman. Category-specific medial temporal lobe activation and the consolidation of semantic memory: evidence from fmri. *Cognitive Brain Research*, 17:484–494, 2003.

D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement, Volume II*. Academic Press, New York NY, 1989.

John Locke. *An Essay Concerning Human Understanding*. William Tegg, London, 1689.

A. López. The diversity principle in the testing of arguments. *Memory & Cognition*, 23(3):374 – 382, 1995.

A. López, S. Atran, J. Coley, D. Medin, and E. Smith. The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32:251–295, 1997.

A. Martin. Functional neuroimaging of semantic memory. In R. Cabeza and A. Kingstone, editors, *Handbook of functional neuroimaging of cognition*, pages 153–186. MIT Press, Cambridge MA, 2001.

A. Martin. The representation of object concepts in the brain. *Annual Review of Psychology*, 58: 25–45, 2007.

A. Martin, C. L. Wiggs, L. G. Ungerleider, and J. V. Haxby. Neural correlates of category specific behavior. *Nature*, 379:649–652, 1996.

D. L. Medin, R. L. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100: 254–278, 1993.

R. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, New York NY, 1990.

D. Osherson, E. E. Smith, and E. Shafir. Some origins of belief. *Cognition*, 24:197–224, 1986.

D. Osherson, E. Smith, O. Wilkie, A. López, and E. Shafir. Category-Based Induction. *Psychological Review*, 97(2):185–200, 1990.

D. Perani, T. Schnur, T. Tettamanti, M. Gorno-Tempini, S. F. Cappa, and F. Fazio. Word and print matching: A pet study of semantic category effects. *Neuropsychologia*, 37:293–306, 1999.

B. Rehder. Property generalization as causal reasoning. In A. Feeney and E. Heit, editors, *Inductive Reasoning: Experimental, Developmental, and Computational Approaches*, pages 81 – 133. Cambridge University Press, Cambridge UK, 2007.

B. Rehder. When similarity and causality compete in category-based property induction. *Memory & Cognition*, 34:3 – 16, 2006.

C. Rorden and M. Brett. Stereotaxic display of brain lesions. *Behavioral Neurology*, 12:191–200, 2000.

E. M. Saffran and M. F. Schwartz. Of cabbages and things: Semantic memory from a neuropsychological perspective — a tutorial review. In C. Umilta and M. Moscovitch, editors, *Attention and Performance*, volume XV, pages 507–536. Churchill Livingstone, Hove and London, 1994.

S. Sloman. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press, Oxford UK, 2005.

S. A. Sloman. Feature based induction. *Cognitive Psychology*, 25:231–280, 1993.

E. E. Smith, E. Shafir, and D. Osherson. Similarity, plausibility, and judgments of probability. *Cognition*, 49:67–96, 1993.

J. B. Tenenbaum, C. Kemp, and P. Shafto. Theory-based bayesian models of inductive reasoning. In A. Feeney and E. Heit, editors, *Inductive Reasoning: Experimental, Developmental, and Computational Approaches*, pages 167 – 2004. Cambridge University Press, Cambridge UK, 2007.

K. Tentori, N. Bonini, and D. Osherson. The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28:467477, 2004.

S. Thompson-Schill, I. P. Kan, and R. T. Oliver. Functional neuroimaging of semantic memory. In R. Cabeza and A. Kingstone, editors, *Handbook of Functional Neuroimaging of Cognition, 2nd Edition*, pages 149–190. MIT Press, Cambridge MA, 2006.

A. Tversky. Features of Similarity. *Psychological Review*, 84:327–352, 1977.

E. K. Warrington and T. Shallice. Category specific semantic impairments. *Brain*, 107:829–854, 1984.

A. Wayne. Bayesianism and Diverse Evidence. *Philosophy of Science*, 62:111 – 121, 1995.

M. Weber, S. Thompson-Schill, D. Osherson, J. Haxby, and L. Parsons. Predicting judged similarity of mammals from their neural representations. in preparation.

M. J. Webster, J. Bachevalier, and L. G. Ungerleider. Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral Cortex*, 4:470–483, 1994.

D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45):1229212307, November 2007.

TABLE 1: Average estimates of the 40 conditional and 9 unconditional probabilities

Argument	*Rated prob*	*Argument*	*Rated prob*
$\mathit{Prob}(\text{ dolphins } \mid \text{ horses })$	.516	$\mathit{Prob}(\text{ bears } \mid \text{ hippos })$	.567
$\mathit{Prob}(\text{ hippos } \mid \text{ elephants })$	.678	$\mathit{Prob}(\text{ camels } \mid \text{ giraffes })$	.677
$\mathit{Prob}(\text{ lions } \mid \text{ cougars })$	.758	$\mathit{Prob}(\text{ lions } \mid \neg\text{camels })$	.396
$\mathit{Prob}(\text{ cougars } \mid \neg\text{horses })$	.421	$\mathit{Prob}(\text{ dolphins } \mid \neg\text{horses })$	.414
$\mathit{Prob}(\text{ giraffes } \mid \neg\text{camels })$	.381	$\mathit{Prob}(\text{ elephants } \mid \neg\text{hippos })$	.383
$\mathit{Prob}(\neg\text{bears } \mid \text{ horses })$	.388	$\mathit{Prob}(\neg\text{dolphins } \mid \text{ elephants })$	.482
$\mathit{Prob}(\neg\text{lions } \mid \text{ cougars })$	.394	$\mathit{Prob}(\neg\text{elephants } \mid \text{ giraffes })$	.401
$\mathit{Prob}(\neg\text{bears } \mid \text{ dolphins })$	.462	$\mathit{Prob}(\neg\text{dolphins } \mid \neg\text{hippos })$	.596
$\mathit{Prob}(\neg\text{horses } \mid \neg\text{bears })$	.559	$\mathit{Prob}(\neg\text{elephants } \mid \neg\text{hippos })$	.691
$\mathit{Prob}(\neg\text{camels } \mid \neg\text{lions })$	.597	$\mathit{Prob}(\neg\text{giraffes } \mid \neg\text{cougars })$	.605
$\mathit{Prob}(\text{ lions } \mid \text{ bears, dolphins })$	.690	$\mathit{Prob}(\text{ camels } \mid \text{ horses, giraffes })$	.714
$\mathit{Prob}(\text{ dolphins } \mid \text{ elephants, hippos })$	.570	$\mathit{Prob}(\text{ cougars } \mid \text{ lions, giraffes })$	.723
$\mathit{Prob}(\text{ elephants } \mid \text{ dolphins, camels })$	.633	$\mathit{Prob}(\text{ camels } \mid \text{ elephants, horses })$	.674
$\mathit{Prob}(\text{ giraffes } \mid \text{ cougars, hippos })$	.611	$\mathit{Prob}(\text{ hippos } \mid \text{ horses, bears })$	.656
$\mathit{Prob}(\text{ bears } \mid \text{ cougars, lions })$	.696	$\mathit{Prob}(\text{ giraffes } \mid \text{ horses, elephants })$	.763
$\mathit{Prob}(\text{ cougars } \mid \text{ lions, }\neg\text{bears })$	.654	$\mathit{Prob}(\text{ elephants } \mid \text{ hippos, }\neg\text{dolphins })$	.662
$\mathit{Prob}(\text{ giraffes } \mid \text{ camels, }\neg\text{hippos })$	.622	$\mathit{Prob}(\text{ camels } \mid \text{ bears, }\neg\text{dolphins })$	.510
$\mathit{Prob}(\text{ horses } \mid \text{ giraffes, }\neg\text{cougars })$	.573	$\mathit{Prob}(\text{ elephants } \mid \text{ hippos, }\neg\text{bears })$	.626
$\mathit{Prob}(\text{ elephants } \mid \text{ lions, }\neg\text{camels })$	.455	$\mathit{Prob}(\text{ lions } \mid \text{ cougars, }\neg\text{horses })$	.680
$\mathit{Prob}(\text{ hippos } \mid \text{ camels, }\neg\text{dolphins })$	.534	$\mathit{Prob}(\text{ horses } \mid \text{ bears, }\neg\text{giraffes })$	.499
$\mathit{Prob}(\text{ horses })$	.583	$\mathit{Prob}(\text{ hippos })$	.563
$\mathit{Prob}(\text{ dolphins })$	.559	$\mathit{Prob}(\text{ bears })$	.588
$\mathit{Prob}(\text{ elephants })$	.601	$\mathit{Prob}(\text{ giraffes })$	.565
$\mathit{Prob}(\text{ camels })$	.550	$\mathit{Prob}(\text{ cougars })$	.564
$\mathit{Prob}(\text{ lions })$	.633		

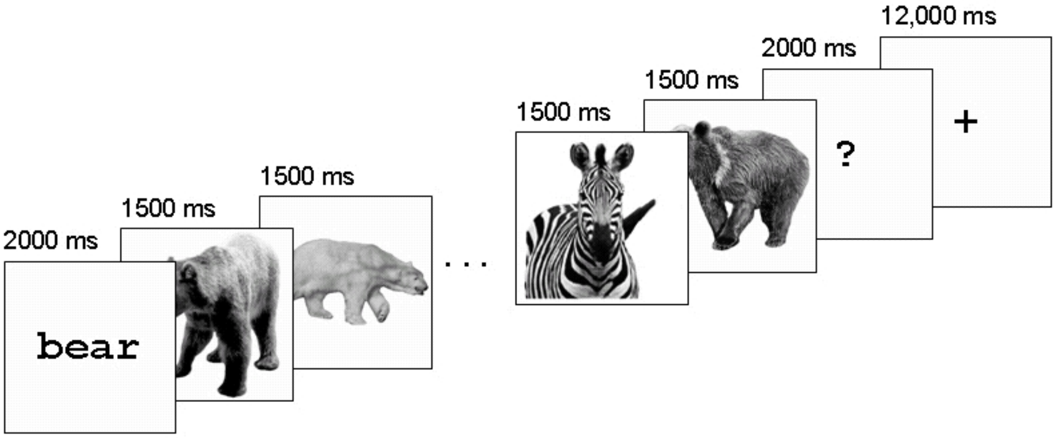
Estimates of probability. The common predicate (suppressed in the table) for all arguments was: *have at least 18% of their cortex in the frontal lobe*. Each number is the average of 20 responses.

TABLE 2: Similarities computed from left ventral-temporal cortex and from left BA 19.

mammals		ventral temporal	BA 19	mammals		ventral temporal	BA 19
lion	cougar	0.650	0.591	hippo	horse	0.608	0.605
hippo	elephant	0.612	0.569	horse	cougar	0.608	0.576
giraffe	camel	0.605	0.641	dolphin	bear	0.409	0.362
giraffe	horse	0.581	0.617	bear	elephant	0.656	0.564
camel	horse	0.628	0.632	horse	lion	0.603	0.666
dolphin	horse	0.365	0.422	camel	cougar	0.659	0.590
lion	bear	0.656	0.628	cougar	giraffe	0.660	0.579
horse	elephant	0.592	0.611	elephant	lion	0.646	0.636
elephant	camel	0.593	0.583	camel	lion	0.630	0.667
hippo	giraffe	0.602	0.622	bear	giraffe	0.636	0.535
bear	cougar	0.658	0.510	camel	bear	0.639	0.596
dolphin	hippo	0.418	0.463	elephant	cougar	0.630	0.500
bear	horse	0.620	0.581	hippo	cougar	0.644	0.557
giraffe	dolphin	0.395	0.401	dolphin	cougar	0.433	0.451
dolphin	elephant	0.333	0.333	dolphin	camel	0.382	0.453
elephant	giraffe	0.593	0.598	lion	hippo	0.667	0.652
camel	hippo	0.649	0.651	lion	giraffe	0.620	0.661
hippo	bear	0.653	0.590	dolphin	lion	0.404	0.467

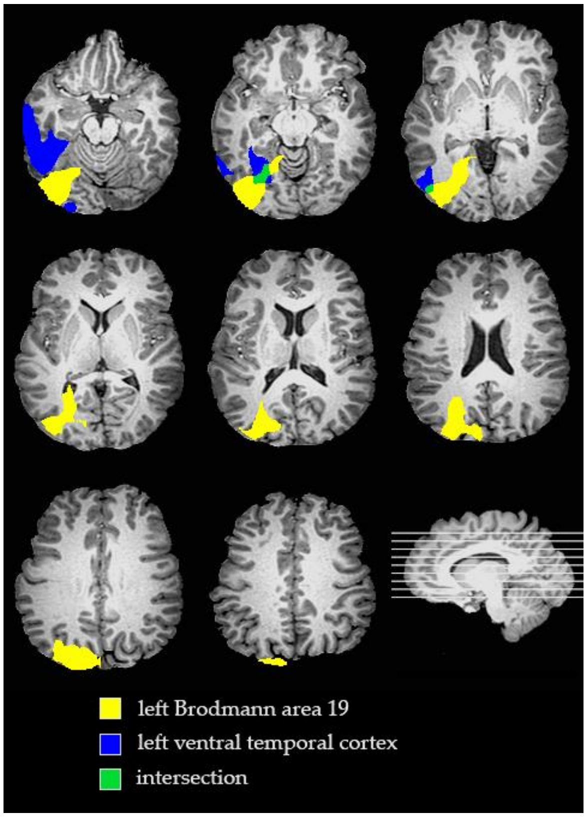
Two estimates of similarities. Each number is derived from the average squared deviation of the two β-coefficients associated with a given voxel when it is "viewing" the respective mammals. The numbers for each area are normalized to the $\left[\frac{1}{3}, \frac{2}{3}\right]$ interval. (See the text for more complete explanation.) The two estimates for the 36 similarities correlate with each other at $r = .840$.

FIGURE 1: fMRI procedure



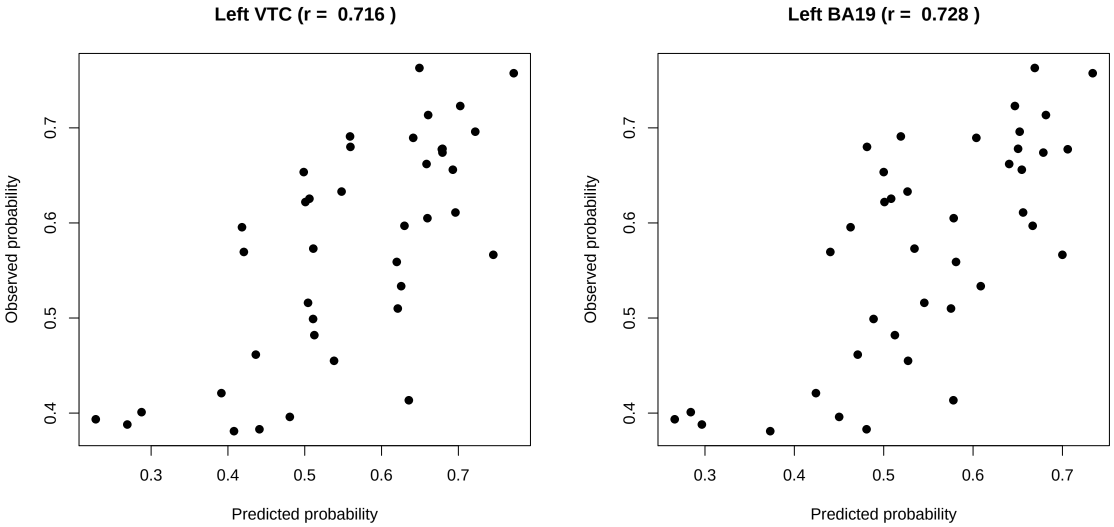
Main task: Spot any animal that is not named at the outset of the trial. In the case pictured here, the zebra must be signaled (a "catch trial"). In most trials there were no intrusions, and brain activations were used only in such non-catch trials. In control trials, phase-scrambled versions of the mammal images were presented, and subjects signaled the presence of a faint cross hatch. Again, these cases served as "catch trials" and were excluded from the analysis.

FIGURE 2: Two regions for extracting similarity from the brain



Horizontal slices for LVTC and LBA19 (positions shown at the bottom right). The two regions were exploited separately to compute the neural similarity of mammal concepts.

FIGURE 3: Scatter plots for results based on two neural regions



Predicted versus observed conditional probabilities: Predictions are obtained using similarities computed from left ventral-temporal cortex and from left BA 19.