From similarity to inference[*]

Matthew Weber
Princeton University

Daniel Osherson
Princeton University

July 28, 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 37 (LBA37) in response to pictures of different categories; the choice of LBA37 is motivated by previous literature. The theory was tested against probability judgments for 160 arguments generated from 16 mammal categories and a common predicate. The theory's predictions (based on neural similarity) correlate strongly with these estimates. Other brain regions previously implicated in semantic cognition yield similarities that also allow the model to predict inductive judgments accurately whereas use of rated similarity in place of neural similarity is less successful.

[*] The authors wish to thank three anonymous referees for constructive commentary. Weber acknowledges support from an NSF graduate research fellowship. Osherson acknowledges the Henry Luce Foundation. Contact: mattheww/osherson@princeton.edu.

1 Introduction

The inference from (1)a to (1)b, below, will likely strike the reader as safer than the one from (1)a to (1)c.

(1) (a) Horses have three ligaments at each knee.

(b) Donkeys have three ligaments at each knee.

(c) Cougars have three ligaments at each knee.

In explanation, it is often observed that donkeys are more similar to horses than cougars are. The explanation offers little insight, however, if similarity is operationalized through numerical ratings of perceived resemblance. The rated resemblance of donkeys and horses, after all, might derive from the conviction that many biological properties true of one apply equally to the other. The appeal to similarity as an account of inference would then be circular.

There is more than one way to operationalize similarity so as to avoid such circularity. To characterize a familiar approach, call a feature of mammals "direct" if the reasoner need not engage in inference to determine whether the feature applies to particular species. For example, *finds much of its food in the ocean*, and *is a carnivore* are direct for most people since the relevant facts are already stored in memory. If similarity is derived from direct features then its role in explaining inferences about non-direct properties [as in (1)] avoids begging the question. Smith et al. (1993), Sloman (1993) and Tenenbaum et al. (2007) implement different versions of this approach.

Note that the use of direct features to predict strength of inference explains the attribution of the target property (e.g., the one about ligaments) in terms of the attribution of other properties, namely, the features. Essentially, confidence that the target property is shared between two categories is presumed to depend on the fraction of known properties that are common to them. Perhaps assessing this fraction exhausts the sense in which similarity participates in inference. Alternatively, inductive inference might rely on a more fundamental source of resemblance involving how categories are mentally represented.

In what follows, we attempt to elucidate the latter kind of resemblance, and exploit it in a theory of inference. For this purpose, we define "mental representation" in physical terms, as the pattern of brain activations provoked by use of a concept. The similarity of two concepts can then be determined by comparing their respective patterns. Concepts are used in many ways but for simplicity we here focus on visual identification of instances. Thus, similarity will be defined by comparing brain activations elicited by classifying images of different category members. This measure is not contaminated by inferential reasoning or deliberate feature evaluation. If successful in predicting probabilities, our approach to mental representation may point the way to a reductive account of inductive reasoning.

Of course, there is more to inference than similarity. For example, the informativeness of premises also counts, as illustrated by the greater strength of the argument from (2)a to (2)b compared to that from (2)c to (2)b.

(2) (a) Whales have fur.

(b) Bats have fur.

(c) Lions have fur.

Similarity must therefore be situated within a theory of inductive inference that embraces additional variables governing the strength of arguments. The usefulness of a given definition of similarity can then be evaluated through the predictive success of the overall model.

But what theory of inductive inference should be used for this purpose? We adopt a probabilistic perspective. Specifically, the conditional probability $\mathit{Prob}\,(\mathsf{A} \mid \mathsf{B}_1 \cdots \mathsf{B}_k)$ is assumed to measure the psychological strength of the inference from $\mathsf{B}_1 \cdots \mathsf{B}_k$ (thought of as premises of an argument) to A (its conclusion). Or rather, this is assumed to be true for a suitable choice of distribution Prob. We'll attempt to construct such a distribution using similarity as an essential ingredient. Our approach thus portrays human judgment as *coherent*, that is, in conformity with the laws of probability. This approach might seem to be undermined by demonstrations of incoherence in ordinary judgment, for example, attributing greater probability to a conjunction than to a constituent.[1] Eliciting such errors, however, requires devious selection of stimuli whereas our materials (described below) seem inoffensive in this regard.[2] We therefore follow Pearl (1988) in conceiving of probability "as a faithful guardian of common sense." If successful to a first approximation, modification of the theory may allow it to illuminate incoherent judgment as well.

The remainder of the paper is organized as follows. The next section reviews elements of subjective probability. Our proposed theory of inference is described in Section 3, and compared to its predecessors. Data on inferences involving mammalian categories are presented subsequently (Section 4), followed by description of our physiologically-based similarity measure (Section 5). The predictive success of the theory is taken up

[1] Such attribution $\mathit{Prob}\,(\mathsf{A} \wedge \mathsf{B}) > \mathit{Prob}\,(\mathsf{B})$ is known as the *conjunction fallacy*. It was first documented by Tversky and Kahneman (1983). For recent research, see Bonini et al. (2004); Tentori et al. (2004); Wedell and Moro (2007) and references cited there.

[2] For one hypothesis about the properties of fallacy-inducing stimuli, see Crupi et al. (2008).

in Section 6, followed by discussion.

The model developed in the next two sections is not a general account of inductive inference. It is adapted to properties of a biological character about which the reasoner has partial information but no access to a causal nexus that determines how the property is distributed across categories. In contrast, much of ordinary reasoning relies on causal schemas (Rehder, 2006, 2007; Sloman, 2005). In fact, our model is easily adapted to represent many forms of causal knowledge as briefly indicated at the end.

2 Subjective probability

Only a few concepts are needed (fuller treatment is available in Jeffrey, 1983; Nilsson, 1986; Halpern, 2003).[3] Let $s_1 \cdots\ s_n$ be n *sentential variables*, that is, statements with a definite truth-value (either true or false). Typically, the s_i describe non-repeatable circumstances whose probability cannot be estimated through sampling. For example, one variable might affirm that horses have three ligaments at each knee, another that this is true of donkeys. The variables generate an infinite set of *formulas* through closure under conjunction ($\bigwedge$), disjunction ($\bigvee$), and negation ($\neg$).

By a *complete conjunction* is meant a conjunction $\pm s_1 \bigwedge \cdots \bigwedge \pm s_n$ of all the sentential variables, each with either positive or negative polarity. The left column in Table 1 displays the eight complete conjunctions that arise from three variables, abbreviated as p, q, r. Any assignment of numbers to the complete conjunctions is called a *(probability) distribution* over the n sentential variables, provided that the numbers are nonnegative and sum to one. An example is provided in the table. A given distribution *Prob* determines the probability of all formulas since each formula is equivalent to a disjunction

[3] An elementary exposition is available at `http://www.princeton.edu/~osherson/primer.pdf`.

of complete conjunctions. For example, $p \wedge q$ is equivalent to the disjunction of the first two complete conjunctions in Table 1, and $p \vee q$ is equivalent to the disjunction of all but the last two.[4] The probability of a formula φ is obtained by summing $\mathit{Prob}(C)$ over all complete conjunctions C in the disjunction to which φ is equivalent (summing is appropriate because distinct complete conjunctions are logically incompatible). For the distribution in the table, $\mathit{Prob}(p \wedge q) = .3$ and $\mathit{Prob}(p \vee q) = .7$.

All *conditional* probabilities are likewise determined via the familiar equation:

$$\mathit{Prob}(A \mid B_1 \cdots B_k) = \frac{\mathit{Prob}(A \wedge B_1 \wedge \cdots \wedge B_k)}{\mathit{Prob}(B_1 \wedge \cdots \wedge B_k)}.$$

To illustrate with the table:

$$\begin{aligned}
\mathit{Prob}(q \mid p, \neg r) &= \frac{\mathit{Prob}(q \wedge p \wedge \neg r)}{\mathit{Prob}(p \wedge \neg r)} \\
&= \frac{\mathit{Prob}(p \wedge q \wedge \neg r)}{\mathit{Prob}((p \wedge q \wedge \neg r) \vee (p \wedge \neg q \wedge \neg r))} \\
&= \frac{\mathit{Prob}(p \wedge q \wedge \neg r)}{\mathit{Prob}(p \wedge q \wedge \neg r) + \mathit{Prob}(p \wedge \neg q \wedge \neg r)} \\
&= \frac{.2}{.2 + .1} = 2/3.
\end{aligned}$$

In the next section we propose a principle for specifying the probability of complete conjunctions, hence, for calculating conditional probabilities. The following fact will be pivotal (for discussion and proof, see Neapolitan, 1990).

(3) FACT: Let conjunction $C = A_1 \wedge \cdots \wedge A_k$ and distribution *Prob* be given. Then $\mathit{Prob}(C)$ falls in the interval with lower bound

$$\max\{0,\ 1 - k + \sum_{i=1}^{k} \mathit{Prob}(A_i)\}$$

[4] That is, $p \wedge q$ is logically equivalent to $(p \wedge q \wedge r) \vee (p \wedge q \wedge \neg r)$.

and upper bound

$$\min\{\mathit{Prob}\,(A_1)\cdots\mathit{Prob}\,(A_k)\}.$$

Moreover, each point in the interval is attained by some choice of *Prob*.

For example, let $C = \mathit{Prob}\,(p \wedge \neg q \wedge r)$ and suppose that $\mathit{Prob}\,(p) = .7$, $\mathit{Prob}\,(\neg q) = .8$, $\mathit{Prob}\,(r) = .9$. Then the lower bound on the probability of C is the greater of 0 and

$$1 - 3 + \mathit{Prob}\,(p) + \mathit{Prob}\,(\neg q) + \mathit{Prob}\,(r) = 1 - 3 + .7 + .8 + .9 = .4.$$

The upper bound is the minimum of $\{\mathit{Prob}\,(p), \mathit{Prob}\,(\neg q), \mathit{Prob}\,(r)\}$, that is, .7. Thus, $\mathit{Prob}\,(C) \in [.4, .7]$. If $\mathit{Prob}\,(r) = .3$ instead of .9 then $\mathit{Prob}\,(C) \in [0, .3]$.

3 A theory of inference

3.1 Starting point

Fix a property Q (like the one about ligaments), and let $b_1 \cdots b_n$ be categories to which Q can be sensibly applied (e.g., mammals). We take the sentential variables to be the n sentences of the form Qb_i, that is, the attribution of Q to b_i. In the intended case the reasoner recognizes the biological character of Q but lacks definite information about its application to the categories b_i. To construct probabilities of complete conjunctions, we start from two kinds of information:

(4) (a) the *similarity* between all pairs of categories b_i, b_j.

(b) $\mathit{Prob}\,(Qb_1) \cdots \mathit{Prob}\,(Qb_n)$, that is, the probability of each variable.

The similarity function will be denoted *sim*$(\cdot,\cdot)$. It is assumed that *sim*$(a,b) \in [0,1]$, *sim*$(a,b) =$ *sim*(b,a), and *sim*$(a,a) = 1$ for all categories a, b.[5]

The theory can now be summarized as follows. Via Fact (3), the probability of a conjunction $Qb_1 \wedge \cdots \wedge Qb_n$ falls in an interval determined by *Prob*$(Qb_1) \cdots$ *Prob*(Qb_n); the point in the interval that corresponds to *Prob*$(Qb_1 \wedge \cdots \wedge Qb_n)$ is chosen as a function of the minimum similarity among $b_1 \cdots b_n$. This scheme is adjusted in a natural way to handle negated conjuncts (as in $Qb_1 \wedge \neg Qb_2 \wedge Qb_3$). Conditional probabilities are derived from the probabilities of conjunctions. We now provide details.

3.2 Conjunctions without negations

To motivate a key idea, consider the strange conjunction $C_0 = Qb \wedge \cdots \wedge Qb$ all of whose conjuncts are identical. Plainly, *Prob*$(C_0) =$ *Prob*(Qb), which is the "minimum" of the probabilities of the conjuncts of C_0, hence equal to the upper bound of the interval for *Prob*(C_0) given by Fact (3). In this case, the similarity between categories in different conjuncts is unity [because *sim*$(b,b) = 1$], hence the *minimum similarity* between categories drawn from different conjuncts is also unity (we justify the use of *minimum* similarity shortly). Generalizing to arbitrary conjunctions $C = Qb_1 \wedge \cdots \wedge Qb_k$, we let $\min\{$*sim*$(b_i, b_j) \mid i, j \leq k\}$ determine the position of *Prob*(C) in the interval defined for C by Fact (3), with greater similarity pushing the probability closer to the upper bound. The simplest expression of this idea relies on a convex sum, as follows.

(5) Probability constructed for $C = Qb_1 \wedge \cdots \wedge Qb_k$:

- Let $\mathbf{p} = \max\{0,\ 1 - k + \sum_{i=1}^{k}$ *Prob*$(Qb_i)\}$, the least possible value of $\mathbf{C}$.

[5] Asymmetry in judged similarity is reported in Tversky (1977) but detected less often in Aguilar and Medin (1999). Symmetry simplifies what follows but is not essential.

- Let $\mathbf{P} = \min\{\mathit{Prob}\,(\mathsf{Qb}_1), \cdots, \mathit{Prob}\,(\mathsf{Qb}_k)\}$, the greatest possible value of $\mathbf{C}$.

- Let $\mathbf{sim} = \min\{\mathit{sim}(\mathsf{b}_i, \mathsf{b}_j) \mid i, j \leq k\}$.

Then we set:

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

To illustrate, let $\mathsf{C} = \mathsf{Qb}_1 \wedge \mathsf{Qb}_2 \wedge \mathsf{Qb}_3$ with:

(6)

$\mathit{Prob}\,(\mathsf{Qb}_1) = .5$	$\mathit{sim}(\mathsf{b}_1, \mathsf{b}_2) = .3$
$\mathit{Prob}\,(\mathsf{Qb}_2) = .8$	$\mathit{sim}(\mathsf{b}_1, \mathsf{b}_3) = .6$
$\mathit{Prob}\,(\mathsf{Qb}_3) = .9$	$\mathit{sim}(\mathsf{b}_2, \mathsf{b}_3) = .7$

Then $\mathbf{p} = .2$, $\mathbf{P} = .5$, $\mathbf{sim} = .3$, and $\mathit{Prob}\,(\mathsf{C}) = [.2 \times (1 - .3] + [.5 \times .3] = .29$.

One motivation for the use of minimum similarity (rather than average or maximum, for example) is that (5) implies the conjunction law:

$$(7)\;\; \mathit{Prob}\,(\mathsf{Qb}_1 \;\&\; \cdots \;\&\; \mathsf{Qb}_n) \;\geq\; \mathit{Prob}\,(\mathsf{Qb}_1 \;\&\; \cdots \;\&\; \mathsf{Qb}_n \;\&\; \mathsf{Qb}_{n+1}).$$

Although not tested overtly, we expect general agreement with the law in the context of the stimuli to be described later. It is easy to see that use of average or maximum may violate (7).

3.3 The general case

It remains to consider conjunctions with negated conjuncts, e.g., $\mathsf{Qb}_1 \wedge \mathsf{Qb}_2 \wedge \neg\mathsf{Qb}_3$. They are treated as before except that one minus the probability of Qb_i is substituted if $\neg\mathsf{Qb}_i$ appears as a conjunct. We also use one minus the similarity of two categories that appear in statements of opposite polarity. For example, the high similarity of horses and donkeys should lower the probability of Q(horse) & $\neg\mathsf{Q}$(donkey) and raise the probability of $\neg\mathsf{Q}$(horse) & $\neg\mathsf{Q}$(donkey). We illustrate with $\mathsf{C} = \mathsf{Qb}_1 \wedge \mathsf{Qb}_2 \wedge \neg\mathsf{Qb}_3$. Suppose that probabilities and similarities are given by (6). Then:

$$\begin{aligned}
\mathbf{p} &= \max\{0, 1 - 3 + .5 + .8 + (1 - .9)\} = \max\{0, -.6\} = 0. \\
\mathbf{P} &= \min\{.5 + .8 + (1 - .9)\} = .1. \\
\mathbf{sim} &= \min\{.3, (1 - .6), (1 - .7)\} = .3. \\
\mathit{Prob}(\mathsf{C}) &= [0 \times (1 - .3)] + [.1 \times .3] = .03.
\end{aligned}$$

The foregoing algorithm constructs probabilities for every conjunction over the set of variables and their negations. Contradictions are assigned zero probability; that is, according to our scheme (as easily demonstrated):

(8) $\mathit{Prob}(\mathsf{Qb}_1 \wedge \ldots \wedge \mathsf{Qb}_i \wedge \ldots \wedge \neg\mathsf{Qb}_i \wedge \ldots \mathsf{Qb}_n) = 0.$

In contrast, (8) fails if average or max is substituted for minimum similarity in the theory.

3.4 Inductive strength of arguments

Although the algorithm maps each complete conjunction to the unit interval, there is no guarantee that these numbers sum to 1. Division of each number by their sum is typically necessary to obtain a probability distribution. For the purposes of computing conditional probabilities, however, the normalization step can be omitted because of the division inherent in calculating conditional probabilities. Thus, for the argument with premises $\mathtt{Qb}_1$, $\neg\mathtt{Qb}_2$, and conclusion $\mathtt{Qb}_3$, it suffices to construct the two probabilities $\mathit{Prob}\,(\mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2 \wedge \mathtt{Qb}_3)$ and $\mathit{Prob}\,(\mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2 \wedge \neg\mathtt{Qb}_3)$. The strength of the argument is then predicted from the quotient

$$\frac{\mathit{Prob}\,(\mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2 \wedge \mathtt{Qb}_3)}{\mathit{Prob}\,(\mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2 \wedge \mathtt{Qb}_3) + \mathit{Prob}\,(\mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2 \wedge \neg\mathtt{Qb}_3)}$$

which defines $\mathit{Prob}\,(\mathtt{Qb}_3 \mid \mathtt{Qb}_1, \neg\mathtt{Qb}_2)$.

It is worth observing that our theory exhibits the *premise diversity effect* in the sense that $\mathit{Prob}\,(\mathtt{Qb}_3 \mid \ \mathtt{Qb}_1, \mathtt{Qb}_2)$ tends to be greater for smaller values of $\mathit{sim}(\mathtt{b}_1, \mathtt{b}_2)$. More precisely, this holds if the similarities of $\mathtt{b}_1$ and $\mathtt{b}_2$ to $\mathtt{b}_3$ are not too high.[6] 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 considered in Osherson et al. (1990); López et al. (1997); López (1995); Choi et al. (1997), and elsewhere.

It should also be observed that different similarity functions may lead our theory to the same high correlation with rated probability. To illustrate, suppose that good predictions require the probability of the conjunction $\mathtt{Qb}_1 \wedge \mathtt{Qb}_2 \wedge \mathtt{Qb}_3$ to be close to its lower boundary. Because of the role of *minimum* in our theory, such a probability is achieved when any of $\mathit{sim}(\mathtt{b}_1, \mathtt{b}_2), \mathit{sim}(\mathtt{b}_1, \mathtt{b}_3), \mathit{sim}(\mathtt{b}_2, \mathtt{b}_3)$ are assigned small values. Of course, the more arguments to be predicted, the greater the overall constraint on the similarity function.

The foregoing model is not intended as a detailed description of mental operations; for example, people are unlikely to evaluate all complete conjunctions over even a mod-

[6] When the latter two similarities approach unity, so must the similarity between $\mathtt{b}_1$ and $\mathtt{b}_2$; for, in the limit, similarity becomes identity, which is transitive. Small values of $\mathit{sim}(\mathtt{b}_1, \mathtt{b}_2)$ are thus difficult to interpret if both $\mathit{sim}(\mathtt{b}_1, \mathtt{b}_3)$ and $\mathit{sim}(\mathtt{b}_2, \mathtt{b}_3)$ are high.

est number of variables. But they may assess the rough plausibility of relevant complete conjunctions, and use the appropriate ratio as a guide to argument strength. Our model is an idealization of one way this process might be structured (via unconditional probability and similarity).

3.5 Comparison to previous theories

Consider the predicate *can distinguish colors by moonlight.* Many people will find it more likely to be true of leopards (nocturnal predators) compared to horses. The *similarity-coverage model* (Osherson et al., 1990) does not integrate this information into its predictions about argument strength since it is based only on similarity and category membership. Indeed, the model is only adapted to "blank" predicates, for which the reasoner makes no distinction between the likelihood of applying to one category rather than another. (For most people, one such predicate is *requires biotin to survive.*) This is a severe limitation inasmuch as the conditional probability of an argument's conclusion given its premises is influenced by the prior, unconditional probabilities of each [see (2), above]. The same limitation characterizes Sloman (1993)'s feature-based theory as well as the evolutionary model of Tenenbaum et al. (2007). The "Gap" models advanced in Smith et al. (1993) and Blok et al. (2007a,b) are sensitive to prior probabilities but do not define a distribution over all complete conjunctions. They hence offer probabilities for a limited set of arguments and may also violate coherence. As seen above, the current theory integrates prior probabilities and defines a distribution. It is closest to the algorithm QPf described in Blok et al. (2007b). But we here avoid a quadratic programming step among other improvements.

The *coverage* variable defined in Osherson et al. (1990) is represented by the probability of $\mathsf{Qb}_1 \wedge \cdots \wedge \mathsf{Qb}_n$ where $\mathsf{b}_1 \cdots \mathsf{b}_n$ is the class of mammals that come readily to the reasoner's mind. It is claimed in Osherson et al. (1990) that coverage affects the strength of unquantified arguments like $\mathsf{Qb}_1, \mathsf{Qb}_2 \,/\mathsf{Qb}_3$. If this is true, coverage as defined above can be integrated into the present theory in several ways (details omitted).

4 Probability data

We collected estimates of probability for arguments involving the following predicate and mammal categories.

PREDICATE: *have at least 18% of cortex in the frontal lobe*

CATEGORIES:

bear	camel	chimpanzee	cougar
deer	dolphin	elephant	fox
giraffe	hippo	horse	lion
panda	rat	squirrel	wolf

Five types of arguments were constructed, as summarized in Table 2. Thirty-two Princeton undergraduates were recruited to attach conditional probabilities to the 160 arguments with premises, and unconditional probabilities to the 16 arguments without. The 16 unconditional probabilities provide the information shown in (4)b, needed to apply our theory; they correspond to the 16 categories listed above. The 160 conditional probabilities allow the theory to be tested. For each of the arguments with premises, the categories figuring in it were chosen randomly under the constraint that no two arguments have the same conclusion and premises.

The 160 arguments with premises were randomly partitioned into four subsets of 40, each composed of ten arguments of each type. Eight participants evaluated the arguments of one subset plus the 16 arguments with no premises (56 arguments in all). Arguments were presented in random order via computer interface; revision of earlier answers was allowed. The idea of subjective probability (conditional and unconditional) was briefly explained prior to collecting data. All responses were constrained to fall in the unit interval (coded as percents). To illustrate, the last example in Table 2 gave rise to the following query.

> What is the probability that camels have at least 18% of their cortex in the frontal lobe assuming that this is true for giraffes but not dolphins?

The eight responses for each argument with premises were averaged, and likewise for the 32 responses for arguments with no premises. Henceforth, these averages are called *the probabilities* (conditional and unconditional) of the corresponding arguments. The mean unconditional probability ($N = 16$) was 0.560 with minimum 0.401, maximum 0.760, and standard deviation 0.096. For conditional probabilities ($N = 160$), the mean was 0.550 with minimum 0.200, maximum 0.849, and standard deviation 0.136. The probabilities of all 176 arguments are available in the supplement to this report.[7]

5 Similarity data

We now consider the $\binom{16}{2} = 120$ similarities needed for our model [as specified in (4)a]. They were defined by comparing neural responses to images of mammals. We explain by first specifying the neural region examined, next describing the procedure used to provoke and measure activations, and then presenting our measure of similarity.

[7] Available as `http://www.princeton.edu/~osherson/supplement1.txt`.

5.1 Neural region

Lesions to the left temporal lobe have been associated in some patients with deficits in knowledge of biological categories like mammals, fruit, and vegetables, while sparing knowledge of artifacts like tools and furniture (Warrington and Shallice, 1984; Saffran and Schwartz, 1994; Capitani et al., 2003). Broadly consistent with these findings, single cell recording from monkey inferior temporal cortex reveals neurons that are responsive to natural categories (although their specificity may depend on size and position in the visual field, among other factors; see Zoccolan et al., 2007). Partially converging information is available from neuroimaging with humans. A review of experiments 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). Consistent findings appear in studies of category-naming; see, for example, 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. Stewart et al. (2001) provide particularly direct implication of the posterior left fusiform gyrus for the recognition of objects. They report that repetitive transcranial magnetic stimulation of the latter region disrupts picture naming but not word reading or color naming. No such effect was associated with stimulation of the posterior right fusiform gyrus.

Some aspects of property verification also recruit LBA37. Kan et al. (2003) report activation in this area when subjects are asked to determine whether one item is a physical part of another (e.g., udder is part of a cow); right BA37 is inactive. Likewise, Simmons et al. (2007) show that verifying colors of objects recruits a region of LBA37 that is localized by its greater response to colored compared to grayscale images. The same region does not activate for the verification of motoric properties (e.g., being typically thrown). Right BA37 does not respond differentially in this way. Thus, activation of LBA37 is not restricted to visual presentation of categories. Indeed, Price et al. (2003) report that anterior regions of the left fusiform gyrus (part of LBA37) are activated when the color or size of fruits and vegetables were retrieved from their names. Similarly, Chao et al. (1999) show that regions of the lateral and medial fusiform gyri are recruited in high-level verbal queries about animals (e.g., whether a given species is forest-dwelling).

At the same time, there are 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). It should also be noted that other structures have been implicated in conceptual knowledge, notably, the left inferior frontal and entorhinal cortices (see Thompson-Schill, 2003, and Eichenbaum et al., 2007, respectively). For another example, the premotor cortex may be involved in categorizing manipulable objects like fruit, tools, and clothing, although such evidence is not univocal (e.g. Martin et al., 1996; Gerlach et al., 2002b; for reviews, see Gainotti, 2000 and Martin, 2007). There has been no similar report for mammal categories, however.

On the basis of the foregoing evidence, we focus on left Brodmann area 37 (LBA37), shown in Figure 1. LBA37 encompasses much of the posterior, ventral temporal lobe including the fusiform gyrus. This region appears more consistently than others in the studies reviewed above. By conforming to the contours of a Brodmann area, we forestall adjustments of our region after the fact (to optimize predictive success of the model). As observed later, substitution of several alternative regions also yields good results.

It will be seen below that our measure of similarity depends on deactivations as well as activations, both weak and strong, throughout LBA37. Our results are thus consistent with "coarse coding" conceptions of category representation in the brain, based on on large regions rather than small foci (Haxby et al., 2001).

5.2 fMRI procedure

Twelve new participants from Princeton University were recruited to perform a classification task during fMRI scanning. On each trial, they viewed a category label for one of the 16 mammals used in the behavioral study on probability estimates. Then they viewed a series of different grayscale images of the designated mammal, terminated by a category-intruder (selected from the same 16 categories). The number of images prior to the intruder was variable; the participant's task was to signal its appearance via button-press. Only data collected after presentation of the category label and before the intruder 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 (#), again signaled by a button press, concluding the trial. Only data preceding the cross hatch were analyzed further. Procedural details are provided in the Appendix. All images used as stimuli are available via `http://www.princeton.edu/~osherson/images.tgz`.

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 LBA37 into roughly $1,500$ 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. Finally, for each mammal, the mean activation to that mammal over all voxels within LBA37 was subtracted from each voxel's response to that mammal ("mean correction"). The resulting distribution of corrected values (obtained from subtracting the control then mean correcting) induces a "map" of activations over LBA37. There is one such map for each mammal.

5.3 Neural similarity

Let LBA37 be 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 *dis*similarity in LBA37 of the respective neural representations of horses and camels. Such squared deviations were computed for each of the 120 pairs of distinct mammals. To convert them from dissimilarities to similarities, we linearly transformed the 120 squared deviations to run from $\frac{1}{3}$ to $\frac{2}{3}$ then subtracted each from 1 (to reflect them around $\frac{1}{2}$). 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). Later it will be seen that our results are robust to expansions of the similarity interval.

5.4 Rated similarity

Subsequent to scanning, the 12 fMRI participants rated all 120 pairs of mammals for "conceptual similarity" on a scale from 0 to 100 (where 0 represents the absence of similarity and 100 represents identity). Pairs were presented in random order via computer interface; earlier values could be modified as the participant proceeded. Each participant's 120 numbers were linearly transformed to the unit interval. The 12 data sets were then averaged.

In what follows, similarities based on the fMRI analysis will be called *neural.* Similarities collected after scanning will be called *rated.* Both data sets are given in the Supplementary Materials (see footnote 7).

5.5 Pixelwise similarity

Each of the 24 images depicting a given mammal in our experiment was a 400×400 grid of grayscale pixels (varying between 0 and 255). Across the 24 images, we computed the average intensity at each pixel to create an average image for a given mammal. We defined the *pixelwise* similarity of a given pair of mammals by calculating the average squared deviation of intensity at each pixel of their respective average images; the resulting 120 numbers (one for each pair of mammals) were normalized to the unit interval then subtracted from 1. The correlation between pixelwise similarity and neural similarity (based on LBA37) was close to zero ($r = 0.06$). Mean-correction of the images, and prior subtraction of the scrambled images for a given mammal yields a correlation between pixelwise and neural similarity of $r = .19$. These results suggest that our measure of neural similarity is not tributary to low-level aspects of the images used in the experiment.

6 Results

6.1 Performance of the model using neural similarity

Application of our theory requires specifying the two inputs listed in (4), namely, unconditional probability and similarity. For the first, we rely on the unconditional probabilities collected from participants in the probability experiment. Neural similarity serves as the second input. On this basis, for each of the 160 arguments with premises (Table 2), we derived a conditional probability from our algorithm. The (Pearson) correlation between predicted and observed probabilities is $r = 0.793$ ($p < 0.001$). A scatter plot is provided in Figure 2.

To isolate the role of similarity in the predictions, we reapplied the algorithm but this time fixing all similarities at 0.5. This is equivalent to assigning each conjunction the midpoint of the interval defined by unconditional probabilities alone (see Section 3). Removing similarity in this way lowers the correlation between the predicted and observed conditional probabilities to $r = 0.531$. The difference in the two results (with versus without similarity) is significant by a test between dependent correlations [$t(157) = 7.738$, $p < 0.001$].[8] For another evaluation, we substituted random numbers drawn uniformly from $[\frac{1}{3}, \frac{2}{3}]$ for neural similarities then reapplied our algorithm using the random similarity function. In $1,000$ trials, the average correlation achieved between predicted and observed conditional probabilities was $r = 0.409$ ($sd = 0.060$). None reached the original value of 0.793 (based on neural similarity).

If similarities are scaled to intervals wider than $[\frac{1}{3}, \frac{2}{3}]$, the correlation between predicted and observed probabilities remains highly significant. For example, it is still

[8] The test is described in Bruning and Kintz (1977, §4.14). In what follows, the same test is used for all comparisons between correlations.

0.793 for $[\frac{1}{4}, \frac{3}{4}]$, and descends to 0.745 for $[\frac{1}{10}, \frac{9}{10}]$).

When the 160 arguments are divided into the four sets of 40 indicated in Table 2, the theory performs well on each. See Table 3 and Figure 2. We performed the kind of random similarity test outlined above for each of the four types of argument. In each case, less than 4% of the random sets of similarity produced correlations as high as those based on neural similarity.

For a benchmark opposite to removing similarity from the theory (by setting it uniformly to 0.5), we attempted to optimize prediction of the probability estimates by treating the entire vector of 120 similarities as a free parameter. Specifically, simulated annealing (SA, van Laarhoven, 1988) was used to incrementally convert a random starting point into a vector that maximizes the correlation produced by our theory. SA requires that "neighbors" of a given vector V be generated. We defined them as the result of replacing the value of a randomly chosen coordinate of V with a random choice from $[\frac{1}{3}, \frac{2}{3}]$. (The components of all vectors in the search were confined to the latter interval, just as before.) The highest correlation obtained with this technique was 0.924, significantly higher than the 0.793 obtained from LBA37 ($p \approx 0$). Repeated use of SA produced several similarity vectors that generated correlations with probability around 0.9. The intercorrelation among these similarity vectors is roughly 0.7 whereas their correlation with the LBA37 similarity-vector is typically not significant. These results might reflect a shortcoming in the similarities extracted from LBA37, which may not be ideal for explaining inductive inference. Alternatively, the probability estimates contain unexplainable noise that was successfully fitted by the optimization procedure.

Finally, we derived neural similarity from LBA37 in each of the 12 subjects' brains individually, and attempted to predict the probability data on a within-subject basis. (Recall that probability estimates were obtained from 32 separate participants.) Ten of the resulting correlations were significantly greater than the 0.531 (no similarity) baseline (test for the difference between dependent correlations, $p < 0.01$ in nine cases, $p < 0.05$ in one).

6.2 Performance of the model using rated similarity

Rated and neural similarity are significantly but modestly correlated at $r = 0.355$ ($p < .001$, $df = 118$).[9] To use rated similarity as input to our theory, we first linearly scaled it to the interval $[\frac{1}{3}, \frac{2}{3}]$, just as for neural similarity. With rated similarity, our algorithm yields a poor fit to the 160 probability judgments. The correlation between predicted and observed values in this case is only $r = 0.284$, significantly less than the 0.531 correlation obtained in the absence of similarity $[t(157) = -5.499, p < 0.001]$. As shown in Table 3, rated similarity allows the algorithm to perform well on the four types of arguments considered separately (correlations range from 0.602 to 0.778). But the regression lines for the four types have different slopes and intercepts, compromising the overall correlation. Linearly scaling rated similarity to an interval wider than $[\frac{1}{3}, \frac{2}{3}]$ systematically worsens the overall correlation. For example, $[\frac{1}{4}, \frac{3}{4}]$ produces $r = 0.187$ for the 160 arguments. Tightening the interval improves the correlation to a ceiling of $r = 0.531$, which corresponds to setting all similarities to .5 thus eliminating them from the model.

To better understand the greater inductive potential of neurological compared to rated similarity, we converted similarities to distances by subtracting them from 1 then submitted each set of 120 distances to multidimensional scaling (via principal coordi-

[9] Recall that the same participants contributed rated and neural similarities. At the individual level, 11 of the 12 subjects produced a positive correlation between their rated similarities and similarity defined from their LBA37 (no averaging). But the median correlation is only $r = 0.10$.

nates analysis as implemented in R; see Gower, 1966). Consider rated similarity first. Euclidean distances in the 2-dimensional solution correlate with the input distances at $r = 0.634$. The 16 mammals are arrayed along two axes that correspond well to size and ferocity (or predatory reputation); the same axes emerge in earlier studies of avian and mammal categories (Rips et al., 1973; Caramazza et al., 1976). The 2-dimensional solution for LBA37-based similarity correlates with input distances at $r = 0.937$, but no recognizable axes emerge. The difference between the scaling solutions might reflect use of a strategy that focuses principally on size and ferocity when explicitly rating similarity; no such strategy would intervene when measuring the neural overlap of concept representations. The latter measure may be sensitive to alternative dimensions (difficult to characterize) that are more useful to inductive judgments involving predicates like ours.

Another difference between rated and neurological similarity is their opposite skew, as shown in Figure 3. Note that skew is invariant through random shuffling of mammal names, and through random reassignment of similarities to pairs of mammals. To assess the potency of negative skew for the predictive success of neural similarity, we randomly permuted the 16 category names after calculating neural similarity then reapplied our algorithm using the scrambled similarity function. In 1,000 such permutations, only 9 reached 0.793. We also permuted the neural similarities, leaving names intact. In 1,000 of these trials, only 12 reached 0.793. These results show that the predictive success of neurological similarity depends on more than the distribution of similarities; their connection to the right pairs of mammals also counts. Yet in both of these procedures, the average correlation after permuting was surprisingly high (just above 0.7). Hence, the negative skew for neurological compared to rated similarity seems to partially account for the greater predictive success of the former. We note that the same kinds of permutation also systematically degrade the performance of rated similarity.

6.3 Alternative definitions of similarity

Temporal lobe structures alternative to LBA37 were also used to successfully define neural similarity (via voxel-wise squared deviation, as above). For example, substituting left Brodmann area 22 (superior temporal gyrus) for LBA37 in our theory produces a correlation of $r = 0.741$ with respect to the 160 probability arguments; this performance is significantly better than the no-similarity baseline of $r = 0.531$ ($p \approx 0$) but also significantly worse than the $r = 0.793$ performance of LBA37 ($p < 0.03$). Many Brodmann areas fall into this category, for example, LBA18 (occipital cortex, $r = 0.739$), RBA11 (orbitofrontal cortex, $r = 0.731$), and RBA21 (middle temporal gyrus, $r = 0.738$). Only three Brodmann areas produce similarities that predict the probability data as well as LBA37. LBA44 (Broca's area, pars opercularis) yields $r = 0.787$, RBA18 (extrastriate visual cortex) produces $r = 0.794$, and LBA34 (anterior entorhinal cortex) yields $r = 0.782$. The similarities extracted from LBA37 and RBA18 are significantly correlated ($r = 0.472$). None of the other similarity-correlations are significant.

Several Brodmann areas do not yield predictive similarities. Left and right BAs 7 and 10, for example, produce correlations with probability that are not reliably different from the 0.531 value that results from removing similarity from our theory by setting it uniformly to 0.5. (BA7 is located in the posterior parietal lobe whereas BA10 is frontal polar. See Monti et al., 2007 for discussion of these regions in deductive reasoning.) When neural similarities are defined via individual brains, none of these four regions produces a correlation reliably greater than the 0.531 baseline in any subject.

Brodmann areas are defined cytoarchitecturally (Gazzaniga et al., 2002). To specify a cortical area on functional grounds, for each voxel in the brain we computed the difference between the voxel's activation in response to the intact mammal pictures versus its response to the scrambled visual controls. Voxels with significant positive differences were then submitted to clustering, with voxels v_1, v_2 assigned to the same cluster if (a) they shared a face or (b) v_1 shared a face with a voxel in a cluster containing v_2. Clusters smaller than 125 voxels (125 μl) in volume were not considered. (Our procedure is standard; see Eger et al., 2008 for one of many examples of its application.) Two large clusters emerged in the left hemisphere. One lies in a ventral occipitotemporal region called the "lateral occipital complex," known to activate more for structured compared to unstructured images (Grill-Spector and Malach, 2004). The other occupies the temporal pole, characterized by Patterson et al. (2007) as supporting amodal representations of semantic categories (on the basis of neuropsychological evidence). When our algorithm is supplied with neural similarity derived from these two regions, the correlations with the 160 probability judgments are 0.694 and 0.759, respectively. These two clusters emerge in all 12 subjects. Similarity defined from left LOC yields a correlation reliably greater than 0.531 in ten individual brains ($p < .05$); in right LOC, this drops to nine.

Finally, alternative measures of neural similarity can be substituted for voxel-wise squared deviation. One such measure is the Pearson correlation between the activations of corresponding voxels in response to a given pair of mammals.[10] When this similarity function is computed in LBA37, normalized to $[\frac{1}{3}, \frac{2}{3}]$ and inserted into our algorithm, the resulting correlation with respect to the 160 arguments is $r = 0.710$.

[10] In more detail, let h_i and c_i be the response to horses and camels, respectively, in voxel v_i. Then the correlation between the vectors $(c_1, c_2 \cdots)$ and $(h_1, h_2 \cdots)$ is a measure of neural similarity between the representations of horses and camels.

It is striking that no region alternative to LBA37 predicts probability more accurately when Pearson correlation is used to measure similarity. In a given brain region, the correlation with probability tends to be lower under this similarity measure compared to squared deviation.[11]

7 Discussion

Our operationalization of neural similarity seems to exclude any role of covert inductive inference or deliberate property extrapolation. It is based rather on the overlap of the processes mediating category identification. Such processes appear to occur at multiple sites in the brain, including Brodmann areas in the occipital and frontal lobes (RBA18 and LBA44), in anterior entorhinal cortex (LBA34), and the ventral visual stream (notably, LBA37). The roles of these areas in semantic cognition are well documented, as discussed below (except for LBA37, already treated in Section 5.1). We first offer a cautionary remark. A given region of the brain may afford information relevant to inductive inference but not participate in such inference (which might be spared if the region were incapacitated). It would thus be interesting to determine whether any of the four regions identified here shows differential metabolic engagement by inductive inference compared to a suitable control task.

[11] For a different source of similarity, we considered algorithms for quantifying distance between "synsets" (synonym classes) in the lexical database *WordNet* (Kilgarriff, 2000). Several such schemes are described in Pedersen et al. (2004). Their application to the 16 mammals failed to produce a similarity function of any predictive value for the 160 estimates of conditional probability. (All correlations fell below 0.531, the result of removing similarity from the theory by setting all similarities to $\frac{1}{2}$.) Likewise, use of "Latent Semantic Analysis" (Landauer et al., 1998) to define similarity among mammals proved unsuccessful.

Thompson-Schill (2003) summarizes several studies implicating the left inferior frontal gyrus (including LBA44) in diverse semantic tasks including verb and color generation, semantic classification, and semantic monitoring. Likewise, a review by Buckner and Tulving (1995) suggests that semantic processing activates mainly BAs 44 and 45, and Smith and Jonides (1999) show the involvement of BA44 in object recognition.[12] Regarding LBA34, Suzuki et al. (1997) demonstrated that some of its cells code the identity of a remembered object, exhibiting object selectivity and maintaining activity throughout a delayed match-to-sample task. Consistent with the latter results, Kreiman et al. (2000) report a population of human entorhinal cells that show category selectivity during both perception and imagery. A review by Eichenbaum et al. (2007) suggests that lateral entorhinal cortex is principally concerned with object identity, receiving projections from the ventral visual stream via perirhinal cortex. As for RBA18 (which is roughly coterminous with V2), its predictive success in our study is likely related to its role in constructing object representations (Gerlach et al., 2002a) and executing visually based inferences (Knauff et al., 2003).

The nature of the information captured by the activations in our experiment remains unresolved. It might reflect evaluation of features associated with particular mammals. In the case of LBA37, given its position in the ventral visual stream, we expect many such features to be sensory in character (like size and color). But as noted in Section 5.1, LBA37 appears also to register high-level properties of mammals, like habitat (Chao et al., 1999). The boundary between perceptual and conceptual properties is in any case not sharp; having hands, for example, seems to be both. (See Gross et al., 1985,

[12] On the other hand, Poldrack et al. (1999) and Dapretto and Bookheimer (1999) suggest that the role of LBA44 is principally phonological or syntactic and that semantic operations are executed in nearby but distinct regions of prefrontal cortex, such as BAs 45 and 47.

for evidence that rhesus monkeys represent hands in the inferior temporal cortex.)

It is also possible that neural similarity as we have computed it expresses category similarity more directly, without mediation by the set of elementary properties known to be true or false of particular mammals (such as “is a carnivore,” “possesses claws,” and so forth). Although the present data do not decide the matter, such feature-free representation would be consistent with the well-known difficulty of specifying categories in terms of features that are neatly coded in language. For example, no boolean combination of such features appears to provide a criterion that coincides with the concept *bear* (e.g., a grizzly that lost the ability and desire to eat meat would remain a bear; see Fodor, 1998 for discussion of the general point). The visual criteria used to identify bears, along with the semantic content of the concept, might thus be too abstract for easy description.

Finally, we 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 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 causal knowledge. Although this thesis is speculative at the present time, it is worth observing that several forms of causal knowledge can be integrated into our model. This is achieved by setting the probability of specific complete conjunctions to zero, or imposing independence constraints on their probabilities.

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 (176 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.

Stimuli

Sixteen mammals were presented in the course of the study, as indicated in Section 4. We collected grayscale images illustrating each mammal, subtending about four degrees of visual angle. For each of the nine target mammals, there were 24 such images (12 were mirror reversals of the other 12). Each functional run of the experiment also employed phase-scrambled versions of the same mammal images. (Phase scrambling preserves only the amplitudes of the Fourier spectrum of an image.) Some of the scrambled images were marked with a small, low-contrast pound (#) sign randomly placed in the image.

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 16 mammals for 2 s, then a series of 24 or fewer distinct, intact images of the named mammal, each presented for 667 ms (totaling up to 16 s of images). Each series was terminated by three distinct images (667 ms each) of a species drawn from one of the 15 remaining categories (thus, a mismatch to the initial label). Participants were instructed to press a key at the first appearance of an "intruder" image (that is, an image mismatched to the initial label). Thus, the participant might see nine bear images followed by three squirrel images, and be required to respond to the first squirrel. Different trials displayed varying numbers of images (0 to 24) before the intruder appeared, equated across trials for the 16 mammals. Only time points between the label and the first intruder were mapped to mammal regressors.

The images in the visual baseline task were phase-scrambled versions of the mammal pictures. Each baseline trial employed scrambled images from one mammal category. The form of the visual baseline task was identical to the main task, except that #### (in lieu of a category label) was presented prior to the images, and participants searched for a low-contrast crosshatch (#) in the sequence instead of a category mismatch. Images with # appeared at the end of each trial in positions corresponding to intruders in the main task. Only time points between the appearance of #### and the first image with # were mapped to visual baseline regressors. The study was organized into 8 runs, each including 12 experimental and 6 baseline trials.

All participants responded accurately on all trials. For eight of the twelve subjects there was no significant difference in response times between intact versus scrambled trials (via t-test); two subjects were significantly faster for the intact trials, the other two were significantly faster for the scrambled trials.

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. In a given voxel, the activation level for mammal m was 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 activation maps for a given mammal (one for each subject) were projected into Talairach space and averaged, leaving us with 16 such maps, one for each mammal. Only voxels present in the intersection of all participants' intracranial masks were considered. (That is, when warped into Talairach space, different participants' brains may occupy slightly different volumes; we discarded voxels that were not occupied by everyone.) 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).

Tables

Table 1: Example of a distribution (*Prob*) over three variables.

complete conjunctions	*Prob*
$p \wedge q \wedge r$	.1
$p \wedge q \wedge \neg r$	.2
$p \wedge \neg q \wedge r$	.05
$p \wedge \neg q \wedge \neg r$	.1
$\neg p \wedge q \wedge r$	.05
$\neg p \wedge q \wedge \neg r$	.2
$\neg p \wedge \neg q \wedge r$	.2
$\neg p \wedge \neg q \wedge \neg r$	.1
	1.0

Table 2: Types of arguments

Type	*Example (predicate implicit)*	*Number constructed*
unconditional (no premise)	*Prob* (cougar)	16
one positive premise	*Prob* (panda | bear)	40
one negative premise	*Prob* (horse | not-deer)	40
two positive premises	*Prob* (squirrel | rat, chimpanzee)	40
two premises, one negative	*Prob* (camel | giraffe, not-dolphin)	40

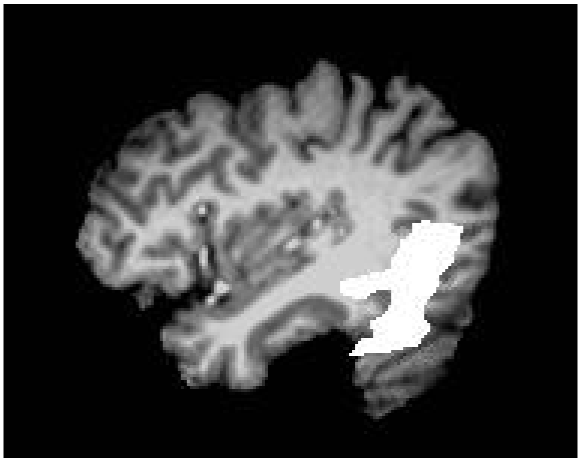
Note: The common predicate (suppressed in the table) for all arguments was: *have at least 18% of their cortex in the frontal lobe*. The categories figuring in each argument were determined randomly.

Table 3: Accuracy using neural vs. rated similarity

Argument type	*number of arguments*	*neural*	*rated*
1 pos. prem.	40	0.771	0.775
1 neg. prem.	40	0.721	0.602
2 pos. prems.	40	0.688	0.718
2 prems, 1 neg.	40	0.703	0.778
overall	160	0.793	0.284

Figures

Figure 1: LBA37, sagittal view



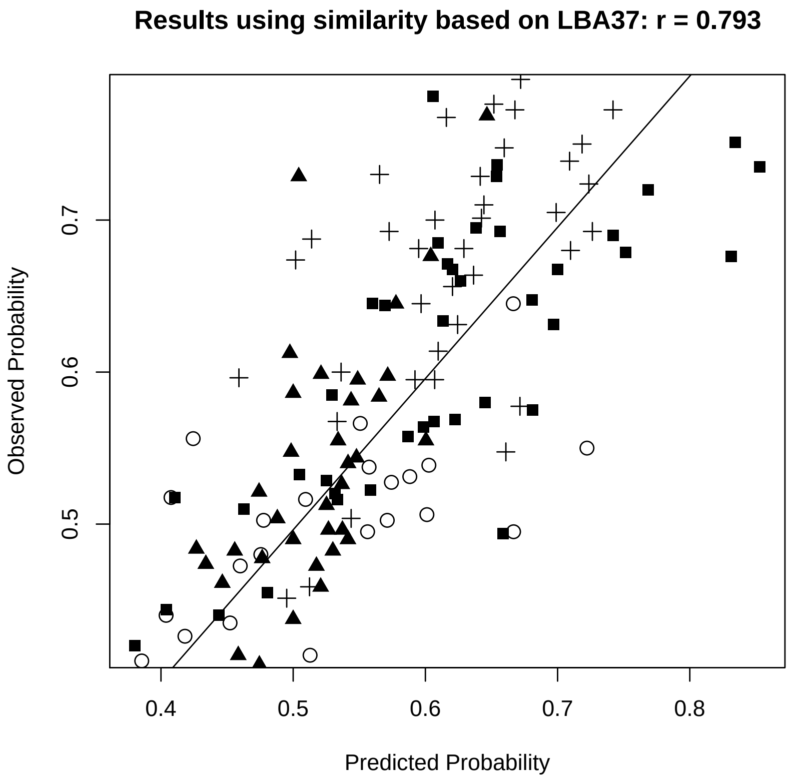


Figure 2: Predicted versus observed probability judgments. Key:

one positive premise	■	one negative premise	○
two positive premises	+	two premises, one negative	▲

Slope of the regression line: 0.632. Intercept: 0.206.

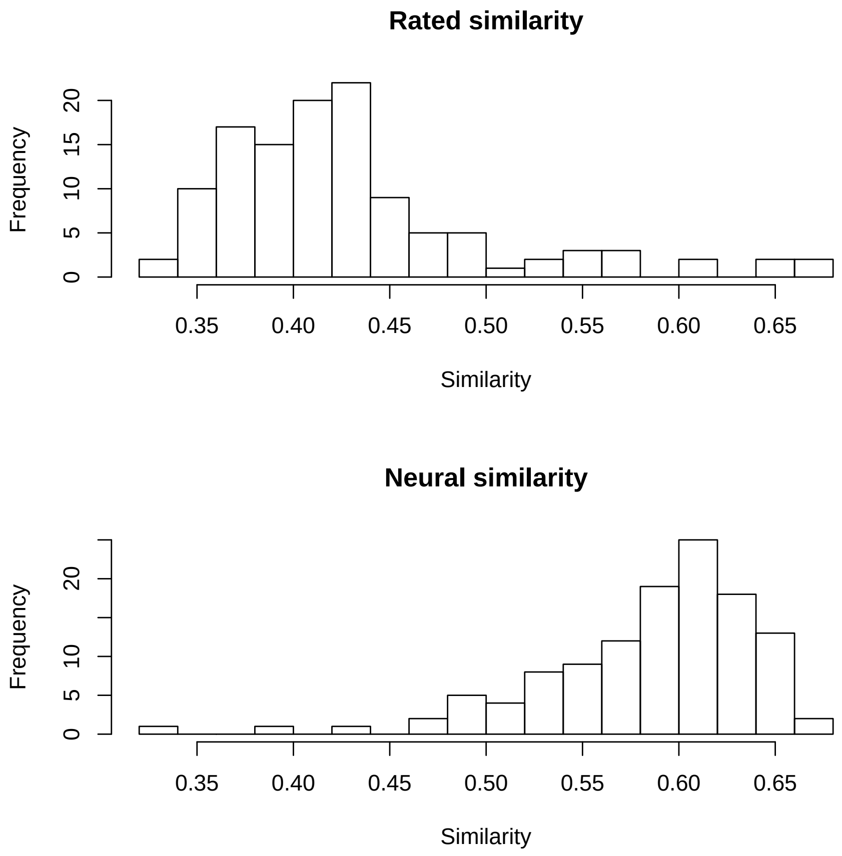


Figure 3: Distributions of rated and neurological similarities

The histograms show the distribution of the 120 similarities (one for each pair of 16 mammals) resulting from rating and from activation patterns in LBA37. In each case, the similarities have been normalized to the interval $[\frac{1}{3}, \frac{2}{3}]$. Notice the difference in skew for the two sources of similarity.

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.

J. L. Bruning and B. L. Kintz. *Computational Handbook of Statistics*. Scott, Foresman and Co., Glenview IL, 2nd edition, 1977.

R. L. Buckner and E. Tulving. Neuroimaging studies of memory: theory and recent pet results. In *Handbook of neuropsychology, vol. 10*. 1995.

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.

A. Caramazza, H. Hersch, and W. Torgerson. Subjective structures and operations in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 15:103–118, 1976.

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.

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

V. Crupi, B. Fitelson, and K. Tentori. Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning*, 2008.

M. Dapretto and S. Y. Bookheimer. Form and content: Dissociating syntax and semantics in sentence comprehension. *Neuron*, 24:427–432, 1999.

E. Eger, J. Ashburner, J.-D. Haynes, R. J. Dolan, and G. Rees. fMRI activity patterns in human LOC carry information about object exemplars within category. *Journal of Cognitive Neuroscience*, 20:356–370, 2008.

H. Eichenbaum, A. P. Yonelinas, and C. Ranganath. The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, 30:123–152, 2007.

J. A. Fodor. *Concepts: Where Cognitive Science Went Wrong.* Clarendon Press, Oxford, 1998.

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.

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.

M. S. Gazzaniga, R. B. Ivry, and G. R. Mangun. *Cognitive Neuroscience: The Biology of the Mind.* W. W. Norton, New York City, 2002.

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

C. Gerlach, I. Law, A. Gade, and O. B. Paulson.

C. Gerlach, C. T. Aaside, G. W. Humphreys, A. Gadeb, O. B. Paulson, and I. Lawa. Brain activity related to integrative processes in visual object recognition: bottom-up integration and the modulatory influence of stored knowledge. *Neuropsychologia*, 40:12541267, 2002a.

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, 2002b.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–328, 1966.

Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu Rev Neurosci*, 27: 649–677, 2004.

C. G. Gross, R. Desimone, T. X. Albright, and E. L. Schwarz. Inferior temporal cortex and pattern recognition. In C. Chagas, R. Gattass, and C. G. Gross, editors, *Pattern Recognition Mechanisms*, pages 179–201. Springer-Verlag, Berlin, 1985.

J. Y. Halpern. *Reasoning about Uncertainty.* MIT Press, Cambridge MA, 2003.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2405–7, 2001.

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

R. C. Jeffrey. *The Logic of Decision (2nd Edition).* The University of Chicago Press, Chicago IL, 1983.

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.

I. P. Kan, L. W. Barsalou, K. O. Solomon, J. K. Minor, and S. L. Thompson-Schill. Role of mental imagery in a property verification task: fmri evidence for perceptual representations of conceptual knowledge. *Cognitive Neuropsychology*, 20(3):525–540, 2003.

A. Kilgarriff. Review of wordnet : An electronic lexical database. *Language*, 76:706–708, 2000.

M. Knauff, T. Fangmeier, C. C. Ruff, and P. N. Johnson-Laird. Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, 15: 559–573, 2003.

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.

G. Kreiman, C. Koch, and I. Fried. Imagery neurons in the human brain. *Nature*, 408:357 – 361, 2000.

T. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

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.

N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–87, 1986.

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

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

K. Patterson, P. J. Nestor, and T. T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8:976–987, December 2007.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pages 1024–1025, San Jose, CA, July 2004.

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.

R. A. Poldrack, A. D. Wagner, M. W. Prull, J. E. Desmond, G. H. Glover, and J. D. E. Gabrieli. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*, 10:15–35, 1999.

C. J. Price, U. Noppeney, J. Phillips, and J. T. Devlin. How is the fusiform gyrus related to category-specificity? *Cognitive Neuropsychology*, 20:561–574, 2003.

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.

L. Rips, E. Shoben, and E. Smith. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12:1–20, 1973.

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.

W. K. Simmons, V. Ramjee, M. S. Beauchamp, K. McRae, A. Martin, and L. W. Barsalou. A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, 45: 2802–2810, 2007.

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 and J. Jonides. Storage and executive processes in the frontal lobes. *Science*, 283:1657–1661, 1999.

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

L. Stewart, B.-U. Meyer, U. Frith, and J. Rothwell. Left posterior BA37 is involved in object recognition: A TMS study. *Neuropsychologia*, 39:1–6, 2001.

W. A. Suzuki, E. K. Miller, and R. Desimone. Object and place memory in the macaque entorhinal cortex. *Journal of Neurophysiology*, 78:1062–1081, 1997.

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*, 28:467 – 477, 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.

S. L. Thompson-Schill. Neuroimaging studies of semantic memory: Inferring 'how' from 'where'. *Neuropsychologia*, 41:280 – 292, 2003.

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

A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315, 1983.

P. van Laarhoven. *Theoretical and computational aspects of simulated annealing*. Center for Mathematics and Computer Science, Amsterdam, 1988.

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

D. H. Wedell and R. Moro. Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 2007.

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):12292 – 12307, November 2007.