Can cognitive processes be inferred from neuroimaging data?

Russell A. Poldrack

Department of Psychology and Brain Research Institute, UCLA, Los Angeles, CA 90095-1563, USA

There is much interest currently in using functional neuroimaging techniques to understand better the nature of cognition. One particular practice that has become common is 'reverse inference', by which the engagement of a particular cognitive process is inferred from the activation of a particular brain region. Such inferences are not deductively valid, but can still provide some information. Using a Bayesian analysis of the BrainMap neuroimaging database, I characterize the amount of additional evidence in favor of the engagement of a cognitive process that can be offered by a reverse inference. Its usefulness is particularly limited by the selectivity of activation in the region of interest. I argue that cognitive neuroscientists should be circumspect in the use of reverse inference, particularly when selectivity of the region in question cannot be established or is known to be weak.

Introduction

Functional neuroimaging techniques such as functional magnetic resonance imaging (fMRI) provide a measure of local brain activity in response to cognitive tasks undertaken during scanning. These data allow the cognitive neuroscientist to infer something about the role of particular brain regions in cognitive function. However, there is increasing use of neuroimaging data to make the opposite inference; that is, to infer the engagement of particular cognitive functions based on activation in particular brain regions. My goal here is to analyze this practice, known as 'reverse inference', and to characterize some limitations on the effectiveness of this strategy. The companion paper in this issue by Henson [1] discusses a complementary strategy for using neuroimaging to distinguish competing cognitive theories.

The goal of cognitive psychology is to understand the underlying mental architecture that supports cognitive functions. To this end, cognitive psychologists examine the effects of task manipulations on behavioral variables, such as response time or accuracy, and use these data to test models of cognitive function. However, it is often not possible to determine on the basis of behavioral variables alone whether a particular cognitive process is engaged, or whether a particular theory of cognitive architecture is correct; for example, there are well-known examples of theoretical indeterminacy based on behavioral data [2]. If

neuroimaging were able to provide information regarding what cognitive processes were engaged in performance of a particular task, cognitive psychologists would have gained a powerful new tool. Researchers outside cognitive psychology are also sometimes interested in using neuroimaging to determine the engagement of particular cognitive processes. For example, philosophers might wish to know the degree to which emotion versus deliberative reasoning plays a role in moral judgments [3].

Inference in neuroimaging

The usual kind of inference that is drawn from neuroimaging data is of the form '*if cognitive process X is engaged, then brain area Z is active*'. Perusal of the discussion sections of a few fMRI articles will quickly reveal, however, an epidemic of reasoning taking the following form:

(1) In the present study, when task comparison A was presented, brain area Z was active.
(2) In other studies, when cognitive process X was putatively engaged, then brain area Z was active.
(3) Thus, the activity of area Z in the present study demonstrates engagement of cognitive process X by task comparison A.

This is a 'reverse inference', in that it reasons backwards from the presence of brain activation to the engagement of a particular cognitive function.

In many cases the use of reverse inference is informal; the presence of unexpected activation in a particular region is explained by reference to other studies that found activation in the same region. However, in some studies the reverse inference is a central feature. In one study [4], subjects were scanned using PET while they performed an economic exchange task in which they had the chance to punish those who defected. Activation was observed in the dorsal striatum when participants subjected defectors to effective punishment; this activation was inferred to reflect the rewarding properties of altruistic punishment. Similarly, a study using fMRI in rats [5] compared activity during pup suckling versus cocaine administration. Greater activity in the dorsal and ventral striatum during suckling compared with cocaine administration led the authors to conclude that 'pup suckling is more rewarding than cocaine' (p. 149). In each of these studies, a cognitive process ('reward') was inferred from activation in a particular brain system (the striatum). Nearly every

Corresponding author: Poldrack, R.A. (poldrack@ucla.edu).
Available online 6 January 2006

Box 1. Bayes' theorem

Bayes' theorem is a result from probability theory that describes how to compute conditional probabilities (the probability of one event *given* some other event). The theorem is generally stated as:

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)} \quad \text{or} \quad P(X|Z) = \frac{P(Z|X)P(X)}{P(Z|X)P(X) + P(Z|\sim X)P(\sim X)}$$

Let's say X and Z are two Bernoulli events (i.e. they either happen or they do not), and that we have some prior belief about the probability of X. Bayes' theorem gives us a way to update our belief given additional evidence, in this case evidence about Z. The quantities in the formula are:

$P(X|Z)$ = the conditional probability of event X given event Z, known as the **posterior probability**

$P(Z|X)$ = the conditional probability of event Z given event X (which is assumed to be known)

$P(X)$ = the probability of event X before any knowledge about Z was obtained, known as the **prior probability** (or simply the **prior**)

$P(Z)$ = the probability of Z regardless of X, known as the **base rate** of Z

As an example, let X represent the occurrence of rain and Z represent the occurrence of clouds in the sky. Let's say that the prior probability of rain on any day (regardless of the presence of clouds) is $P(X)$ = 0.2, the base rate of clouds in the sky is $P(Z)$ = 0.3, and the conditional probability of clouds given the presence of rain is $P(X|Z)$ = 1.0. With these values, the posterior probability of rain *given the presence of clouds* is $P(Z|X)$ = 0.67 according to Bayes' theorem.

Additional discussion of Bayes' theorem can be found at the Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/entries/bayes-theorem/) and Wikipedia (http://en.wikipedia.org/wiki/Bayes_theorem/).

---

neuroimaging paper (this authors' included) uses similar reverse inferences to explain the occurrence of unpredicted regions of activation (for an early discussion of reverse inference, see D'Esposito *et al.* [6]).

It is crucial to note that this kind of 'reverse inference' is not deductively valid, but rather reflects the logical fallacy of affirming the consequent [7]. The syllogism could be made deductively valid if statement (2) were exclusive, such that area Z was active *if and only if* cognitive process X is engaged. However, cognitive neuroscience is generally interested in a mechanistic understanding of the neural processes that support cognition rather than the formulation of deductive laws. To this end, reverse inference might be useful in the discovery of interesting new facts about the underlying mechanisms. Indeed, philosophers have argued that this kind of reasoning (termed 'abductive inference' by Pierce [8]), is an essential tool for scientific discovery [9].

To better understand the information that is provided by reverse inference, it is useful to restate the inference in probabilistic terms [10], in which case the relevant quantities can be determined using Bayes' theorem (see Box 1 for more on Bayes' theorem):

$$P(COG_X|ACT_Z)$$

$$= \frac{P(ACT_Z|COG_X)P(COG_X)}{P(ACT_Z|COG_X)P(COG_X) + P(ACT_Z|\sim COG_X)P(\sim COG_X)}$$

where COG_X refers to the engagement of cognitive process X and ACT_Z refers to activation in region Z. (It should be noted that the prior $P(COG_X)$ is always conditioned on the particular task being used, and should more properly be termed $P(COG_X|TASKy)$; however, for the purposes of simplicity I have omitted this additional conditionalization). This expression uses an expanded form of Bayes' rule that makes clearer the relation to each of the quantities in Table 1. Viewing the reverse inference problem in this way highlights the fact that the degree of belief in a reverse inference depends upon the selectivity of the neural response (i.e. the ratio of process-specific activation to the overall likelihood of activation in that area across all tasks)

as well as the prior belief in the engagement of cognitive process X given the task manipulation $[P(COG_X)]$. This can be seen more clearly when inference is characterized as a probabilistic graph (see Box 2), in which the propagation of uncertainty between levels of inference is made explicit. More generally, this probabilistic approach allows us to characterize the factors that affect the quality of reverse inferences.

Estimating selectivity using the BrainMap database

The greatest determinant of the strength of a reverse inference is the degree to which the region of interest is selectively activated by the cognitive process of interest. If a region is activated by a large number of cognitive processes, then activation in that region provides relatively weak evidence of the engagement of the cognitive process; conversely, if the region is activated relatively selectively by the specific process of interest, then one can infer with substantial confidence that the process is engaged given activation in the region.

It is unfortunately quite difficult to determine clearly the selectivity of activation in a particular brain region. One possible way to estimate selectivity is to use one of the several databases of imaging results currently accessible on the Internet. By searching for studies that show activation in a particular location, one could potentially formulate an estimate of the selectivity of activation in that region. To examine this idea, I used the BrainMap database (http://www.brainmap.org/) [11], which (as of September 2005) contained data from 3222 experimental comparisons in 749 published papers. Although this represents only a portion of the entire neuroimaging literature, the database provides a broad enough sample of different studies to provide a useful proof of concept. I examined the reverse inference that activation in 'Broca's area' implies engagement of language function. As a seed

Table 1. Frequency table for BrainMap database search, showing the number of experimental comparisons identified for each search[a]

	Language study	Not language study
Activated	166	199
Not activated	703	2154

[a]Location of the ROI was [−37,18,18] in Talairach space, extending 10 mm in each direction.

Box 2. Inference as a probabilistic graph

The relationships between experimental manipulations, cognitive processes and observed variables can be expressed as a probabilistic graph (or Bayesian network) (e.g. see [21]) (Figure I). In such a graph, the nodes represent entities and the edges represent conditional probabilities. This graph highlights several important features of reverse inference. First, it makes clear that the cognitive processes (which are of interest in the reverse inference) are conditioned on the particular task manipulation, such that the prior on the cognitive process takes into account the particular task

being performed. Second, it highlights the fact that the strength of the reverse inference depends upon the degree to which the edge of interest is substantially stronger than all other edges leading to the same activation; in the limit, if all other edges have probabilities of zero, the mapping between cognitive process and fMRI activation is 'one-to-one' [22]. Third, it reminds us that fMRI data are not alone in suffering from the reverse inference problem: Reverse inference based on *any* observable data (e.g. behavioral data) is limited by the same characteristics.
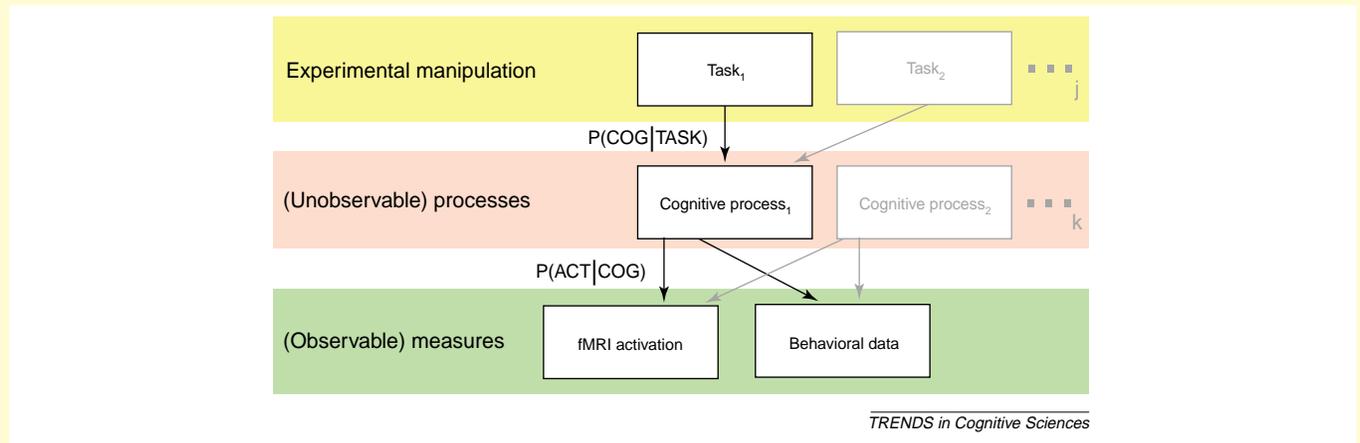

TRENDS in Cognitive Sciences

Figure I. A probabilistic graph representing the relationships between experimental manipulations, cognitive processes and observed variables (see text for details).

Box 3. Region size and selectivity

An interesting question regarding selectivity is how the size of the region being analyzed affects the estimated selectivity of the response. To examine this, searches were performed with four cubes (of widths 4 mm, 12 mm, 20 mm, and 40 mm) centered on the same point used in the analysis in Table 1 in the main text. In Figure II, the posterior probability is plotted against the prior, to demonstrate how the region size affect selectivity. The distance of this function from the diagonal expresses the degree to which the reverse inference provides additional information over the prior, and is proportional to the Bayes factor. These results show that smaller regions are more

selective than large regions, and thus that the power of reverse inference can be maximized by using smaller regions of interest.

It is also useful to ask how these region sizes relate to the kinds of structures that are commonly used in reverse inference. For example, 'Broca's area' is often equated with the left inferior frontal gyrus, pars triangularis. In the AAL atlas [23], this region has a volume of 20104 mm^3, which is roughly equivalent in volume to a 28 mm cube. This suggests that selectivity of many common reverse inferences is likely to be relatively low because of the size of the region.
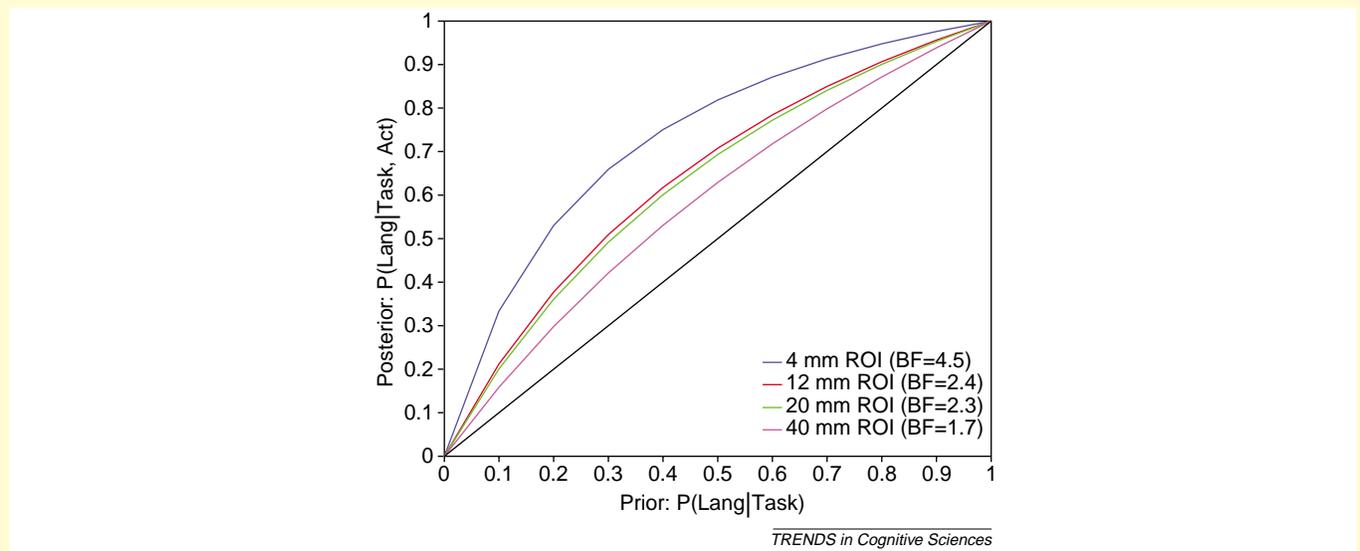

TRENDS in Cognitive Sciences

Figure II. An illustration of how the size of a region of interest (ROI) affects selectivity of the response. Smaller regions give a higher posterior probability (plotted on the ordinate), and hence provide more additional information over the prior than larger regions do.

location, I used a point in the dorsal left inferior frontal gyrus (approximately Brodmann's area 44), which was identified by McDermott *et al.* [12] as being active during engagement of both phonological and semantic processing; a region of interest was created by centering a cube of 20 mm width at this location (for a discussion of the effects of the size of this region, see Box 3). Two searches were performed, one for all experimental comparisons relevant to language processing with activations in this region, and another for all comparisons that were not relevant to language processing (as noted by the Behavioral Domain code in the database) with activation in this region. In addition, the same searches were performed without the anatomical specification, to determine the overall frequency of those classes of studies. The results of these searches are presented in Table 1.

Armed with the results of these searches, one can compute the posterior probability for the reverse inference, which maps onto subjective confidence in that inference. This posterior probability depends both upon the conditional probabilities expressed in Table 1 as well as the prior estimate of language processes being engaged given the particular task. Bayes' rule can be understood as a means of updating one's prior beliefs based on new evidence; positive evidence increases one's belief compared with before, and the degree of this increase depends upon the selectivity of the evidence. Reverse inference will generally be used when we want to infer the presence of a cognitive process that is not directly manipulated by the task, and in this case the prior on engagement of the cognitive process will be relatively low, compared with the case where the process is directly manipulated by the task. If, for example, the prior is 0.5 (i.e. we are equally confident that the process is either engaged or not), then the posterior probability on the inference is 0.69; in other words, activation in the area of interest increases the odds of engagement of the cognitive process from even (1/1) to positive (2.3/1).

How should one determine whether this increase is substantial? One approach is to use the Bayes factor, which is the ratio of the posterior odds to the prior odds, where odds are computed as p/(1–p) [13]. Within the Bayesian inference community, there is a convention attributed to Jeffreys [14] that a Bayes factor between 1 and 3 represents weak evidence, between 3 and 10 reflects moderate evidence, and greater than 10 reflects strong evidence; this is somewhat akin to the convention in frequentist statistics regarding $p < 0.05$. The Bayes factor for the reverse inference discussed above is 2.3, meaning that the inference provides a positive but relatively weak increase in confidence over the prior.

The need for a cognitive ontology

There are several limitations on the foregoing analysis. Most importantly, reverse inference is generally intended to identify the engagement of particular cognitive processes, but this requires that experiments in the database be coded with regard to these cognitive processes. In the language of informatics, this could be termed the 'cognitive ontology' of the database [15]. Unfortunately, the cognitive ontologies of existing
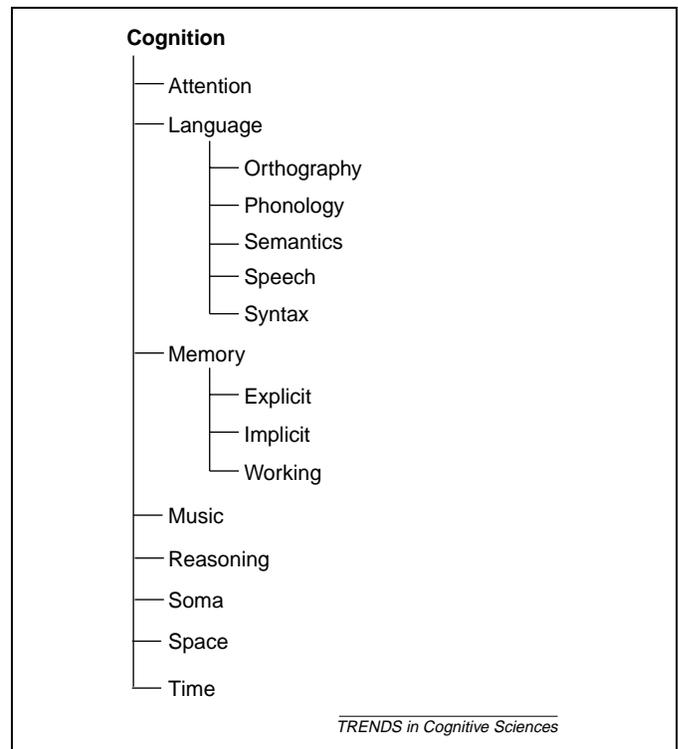
Figure 1. Behavioral domain taxonomy for cognition, from the BrainMap database [11]. In addition to this taxonomy, there are additional taxonomies for Action, Emotion, Interoception, Perception and Pharmacology.

databases are quite coarse in comparison with current theories of cognitive psychology. For example, Figure 1 lists the behavioral domain coding schemes from the BrainMap database for the domain of Cognition. Likewise, the cognitive ontologies used in other databases such as Brede [16] (http://hendrix.imm.dtu.dk/services/jerne/brede/) and the fMRI Data Center [17] (http://fmridc.org) are similarly coarse. Given that these coarse categories are unlikely to map the organization of the mind very cleanly, it seems that powerful reverse inference awaits the development of a detailed cognitive ontology, which will probably require the work of a consortium of cognitive scientists akin to the Gene Ontology consortium (http://www.geneontology.org) that has developed ontologies for genome informatics [18].

Improving reverse inferences

There are two ways in which to improve confidence in reverse inferences: increase the selectivity of response in the brain region of interest, or increase the prior probability of the cognitive process in question. Selectivity is outside the control of the experimenter, but the analysis above suggests that an estimate of selectivity can at least be obtained. In addition, the analysis of sets of regions (functioning as connected networks) might provide greater selectivity than the analysis of single regions, to the degree that specific processes engage specific networks [19]. The size of the region of interest will also affect selectivity (see Box 3), suggesting that reverse inference to smaller regions will provide more confidence.

The prior is to some degree under the control of the experimenter, as he/she can often choose experimental tasks that maximize the prior probability of a particular

Box 4. Questions for future research

- How do brain regions differ in their selectivity?
- Are networks more selective than individual regions?
- Can cognitive psychology support a detailed formal ontology of cognitive proceses?
- How are selectivity estimates from neuroimaging databases biased by selection biases on database entries?

process being engaged. This strategy is more applicable to studies that are directed at making a specific reverse inference, rather than for studies where reverse inference reflects a post hoc explanation for a particular result [20]. One way to increase the prior is by using converging behavioral evidence to provide stronger evidence of engagement of the process of interest. For example, Greene *et al.* scanned subjects using fMRI while they entertained either personal or impersonal moral dilemmas, which were proposed by the authors to differ in the degree to which they engaged emotion in the subjects. Differences in the engagement of several brain regions (medial prefrontal, posterior cingulate, and angular gyrus) were used to infer 'systematic variations in the engagement of emotion in moral judgment' ([3], p. 2107). In parallel to these fMRI results, the investigators also examined response times for trials on which subjects responded that the behavior in question in the dilemma (e.g. pushing a person off a bridge to save several other people) was either appropriate or inappropriate. They found that response times for personal dilemmas were longer when the subjects responded 'appropriate' than when they responded 'inappropriate', whereas the opposite pattern was observed for impersonal dilemmas. They argued that this behavioral effect reflected emotional conflict for the personal but not the impersonal dilemmas, and thus provided converging evidence for the reverse inference. To the degree that such claims regarding the behavioral data are plausible, such a combination of behavioral and fMRI results provides stronger evidence in favor of a reverse inference.

Conclusions

There is substantial excitement about the ability of functional neuroimaging to help researchers to discover the organization of cognitive functions. The analysis presented here suggests that caution should be exercised in the use of reverse inference, particularly in cases where the prior belief in the engagement of a cognitive process and selectivity of activation in the region of interest are low. The results also suggest that mining of neuroimaging databases can provide additional insight into the strength of specific inferences from neuroimaging data, but that the usefulness of these databases is limited by the coarseness of the underlying cognitive ontology used in current databases (see also Box 4). In my opinion, reverse inference should be viewed as another tool (albeit an imperfect one) with which to advance our understanding of the mind and brain. In particular, reverse inferences can suggest novel hypotheses that can then be tested in subsequent experiments. The analysis presented here suggests that this might well be true, but the ultimate usefulness of the reverse inference strategy will be determined by its success in advancing our understanding of the mind and brain in the future.

References

1 Henson, R. (2006) Forward inference using functional neuroimaging: dissociations versus associations. *Trends Cogn. Sci.* 10, doi:10.1016/j.tics.2005.12.005
2 Townsend, J.T. and Ashby, F.G. (1983) *The Stochastic Modeling of Elementary Psychological Processes*, Cambridge University Press
3 Greene, J.D. *et al.* (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108
4 de Quervain, D.J. *et al.* (2004) The neural basis of altruistic punishment. *Science* 305, 1254–1258
5 Ferris, C.F. *et al.* (2005) Pup suckling is more rewarding than cocaine: evidence from functional magnetic resonance imaging and three-dimensional computational analysis. *J. Neurosci.* 25, 149–156
6 D'Esposito, M. *et al.* (1998) Human prefrontal cortex is not specific for working memory: a functional MRI study. *Neuroimage* 8, 274–282
7 Aguirre, G.K. (2003) Functional imaging in behavioral neurology and cognitive neuropsychology. In *Behavioral Neurology and Cognitive Neuropsychology* (Feinberg, T.E. and Farah, M.J., eds), pp. 35–46, McGraw-Hill
8 Pierce, C.S. (1903/1955) Abduction and induction. In *Philosophical Writings of Pierce* (Buchler, J., ed.), pp. 150–156, Dover Books
9 Polya, G. (1954) *Mathematics and Plausible Reasoning*, Princeton University Press
10 Sarter, M. *et al.* (1996) Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *Am. Psychol.* 51, 13–21
11 Fox, P.T. *et al.* (2005) BrainMap taxonomy of experimental design: description and evaluation. *Hum. Brain Mapp.* 25, 185–198
12 McDermott, K.B. *et al.* (2003) A procedure for identifying regions preferentially activated by attention to semantic and phonological relations using functional magnetic resonance imaging. *Neuropsychologia* 41, 293–303
13 Goodman, S.N. (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* 130, 1005–1013
14 Jeffreys, H. (1961) *Theory of Probability*, Clarendon Press
15 Price, C.J. and Friston, K.J. (2005) Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275
16 Nielsen, F.A. *et al.* (2004) Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* 2, 369–380
17 Van Horn, J.D. *et al.* (2004) Sharing neuroimaging studies of human cognition. *Nat. Neurosci.* 7, 473–481
18 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology (The Gene Ontology Consortium). *Nat. Genet.* 25, 25–29
19 McIntosh, A.R. (2004) Contexts and catalysts: a resolution of the localization and integration of function in the brain. *Neuroinformatics* 2, 175–182
20 Poldrack, R.A. and Wagner, A.D. (2004) What can neuroimaging tell us about the mind? Insights from prefrontal cortex. *Curr. Dir. Psychol. Sci.* 13, 177–181
21 Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Patterns of Plausible Inference*, Morgan Kaufmann
22 Henson, R. (2005) What can functional neuroimaging tell the experimental psychologist? *Q. J. Exp. Psychol. A* 58, 193–233
23 Tzourio-Mazoyer, N. *et al.* (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289