A Dynamic Causal Modeling Study on Category Effects: Bottom–Up or Top–Down Mediation?

Andrea Mechelli, Cathy J. Price, Uta Noppeney, and Karl J. Friston

Abstract

■ In this study, we combined functional magnetic resonance imaging (fMRI) and dynamic causal modeling (DCM) to investigate whether object category effects in the occipital and temporal cortex are mediated by inputs from early visual cortex or parietal regions. Resolving this issue may provide anatomical constraints on theories of category specificity—which make different assumptions about the underlying neurophysiology. The data were acquired by Ishai, Ungerleider, Martin, Schouten, and Haxby (1999, 2000) and provided by the National fMRI Data Center (http://www.fmridc.org). The original authors used a conventional analysis to estimate differential effects in the occipital and temporal cortex in response to pictures of chairs, faces, and houses. We extended this approach by estimating neuronal interactions that mediate category effects using DCM. DCM uses a Bayesian framework to estimate and make inferences about the influence that one region exerts over another and how this is affected by experimental changes. DCM differs from previous approaches to brain connectivity, such as multivariate autoregressive models and structural equation modeling, as it assumes that the observed hemodynamic responses are driven by experimental changes rather than endogenous noise. DCM therefore brings the analysis of brain connectivity much closer to the analysis of regionally specific effects usually applied to functional imaging data. We used DCM to estimate the influence that V3 and the superior/inferior parietal cortex exerted over category-responsive regions and how this was affected by the presentation of houses, faces, and chairs. We found that category effects in occipital and temporal cortex were mediated by inputs from early visual cortex. In contrast, the connectivity from the superior/inferior parietal area to the category-responsive areas was unaffected by the presentation of chairs, faces, or houses. These findings indicate that category effects in the occipital and temporal cortex can be mediated by bottom–up mechanisms—a finding that needs to be embraced by models of category specificity. ■

INTRODUCTION

In this study, we combined functional magnetic resonance imaging (fMRI) and dynamic causal modeling (DCM) to investigate neuronal interactions that mediate the representation of objects in the human brain. In recent years, several functional imaging studies have shown that different categories of objects activate a distributed system that includes bilateral fusiform, mid-occipital, and inferior temporal regions. However, within this network, there are areas that respond preferentially to houses, chairs, faces, tools, vehicles, animals, and fruit (e.g., see Haxby et al., 2001; Chao, Haxby, & Martin, 1999; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999, 2000; Cappa, Perani, Schnur, Tettamanti, & Fazio, 1998). These differential effects appear to be relatively small; however, they have been replicated using a number of cognitive tasks (e.g., passive viewing, naming, and semantic decision) and presentation formats (e.g., visual words, photographs, and line drawings). Furthermore, the finding that different regions in the occipital and temporal cortex show preferential responses to different object categories is consistent with reports of category-specific deficits in brain-damaged patients.

A number of hypotheses have been proposed to explain category effects (see Devlin, Russell, et al., 2002, for a review). For instance, Caramazza and Shelton (1998) suggested that distinct regions may be responsible for evolutionarily important categories, such as animals, plants, and tools. In contrast, several authors have proposed that different categories are associated with different types of information—which in turn may lead to functional specialization. For instance, Warrington and Shallice (1984) suggested that perceptual information (i.e., what an object looks like) is more relevant to living objects whereas functional information (i.e., how an object is used) is more relevant for manmade items. Finally, Tyler, Moss, Durrant-Peatfield, and Levy (2000) have proposed that damage to a unified semantic system, undifferentiated by categories or types of information, can still result in deficits that are specific to one or more object categories. In short, the nature of the category effects observed in functional neuroimaging and brain-damaged patients is still debated.

In the present study, we investigate category effects further by looking at the neuronal interactions that

Institute of Neurology, London

mediate the representation of objects in the human brain. Specifically, we combine fMRI and DCM to test whether the category effects observed in the occipital and temporal cortex are mediated by bottom–up or top–down mechanisms. This issue may shape anatomical constraints on theories of category specificity—which make different assumptions about the underlying neurophysiology.

Functional Magnetic Resonance Imaging Data

The data set we used was originally acquired by Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000) and was provided by the National fMRI Data Center (http://www.fmridc.org). In this study, six subjects performed passive viewing and delayed match-to-sample tasks on gray-scale photographs of houses, faces, and chairs. In the passive viewing task, subjects were presented with a series of single stimuli. In the delayed matching task, subjects were presented with a single-sample stimulus followed by a pair of choice stimuli and were asked to indicate which choice stimulus matched the sample stimulus by pressing a button. The baseline for both tasks involved scrambled pictures of houses, faces, and chairs. The authors found that different categories of objects activated a distributed system that included bilateral fusiform, inferior occipital, midoccipital, and inferior temporal regions. However, within this network, there were distinct regions in the occipital and temporal cortex that responded preferentially to faces, house, and chairs (see Figure 1 and Table 2 in Ishai, Ungerleider, Martin, et al., 2000, for details). In short, Ishai et al. analyzed the data using classical inference to estimate the differential effects of houses, faces, and chairs in the occipital and temporal cortex. Here we extend this approach by estimating neuronal interactions in a Bayesian framework as implemented in DCM. In the remaining part of the Introduction, we present DCM briefly and then focus on its application to the data set.

Dynamic Causal Modeling

The aim of DCM is to estimate and make inferences about the influence that one neural system exerts over another and how this is affected by the experimental context. The central ideal is to treat the brain as a dynamic input–state–output system. The inputs correspond to conventional stimulus functions that encode the experimental manipulation. The state variables comprise mean synaptic activities and other biophysical variables that determinate the outputs. The outputs are the regional hemodynamic responses that are measured using fMRI. In DCM, an experiment is regarded as a designed perturbation of neuronal dynamics that is propagated throughout a network of interconnected anatomical nodes. The coupling between regions is therefore estimated by perturbing the system using a series of inputs and then measuring the changes in regionally specific hemodynamic responses. This differs from conventional approaches to brain connectivity, such as multivariate autoregressive models and structural equation modeling. In these models, there is no designed perturbation as it is assumed that the observed hemodynamic responses are driven by endogenous or intrinsic noise.

In DCM, a reasonably realistic but simple neuronal model of interacting neural regions is constructed. This model is then supplemented with an hemodynamic model of fMRI measurements that describes how synaptic activity is transformed into a hemodynamic response (see Mechelli, Price, & Friston, 2001; Friston et al., 2000, for details). The coupling parameters of the neuronal model can thus be estimated from the measured hemodynamic responses. In contrast, existing approaches to brain connectivity usually make inferences by considering the statistical dependencies among hemodynamic responses. However, interactions in the brain occur at the synaptic level, not at the hemodynamic level. DCM accommodates this by including hidden (i.e., unobserved) neuronal and biophysical states when modeling the observed data.

In DCM, three distinct sets of parameters are estimated. A first set of parameters scales the direct and extrinsic influence of inputs on brain states in any particular region. These parameters are generally of little interest in the context of DCM, but, of course, are the primary focus in classical analyses of regionally specific effects. A second set of parameters refers to the intrinsic connections that couple neuronal states in different regions. These parameters allow one to estimate the impact that one neural system exerts over another in the absence of experimental perturbations. A third set of parameters, or "bilinear terms," refer to changes in the intrinsic coupling between regions that are induced by experimental manipulation. These parameters allow one to claim that an experimental manipulation has activated a "pathway" as opposed to a cortical region. By using bilinear terms, DCM accommodates some important nonlinear and dynamic aspects of neuronal interactions. In contrast, multivariate autoregression models and their spectral equivalents, such as coherence analysis, are restricted to linear interactions. Structural equation modeling also assumes that the interactions are linear and, furthermore, instantaneous.

Because dynamic causal models are not restricted to linear or instantaneous systems, they need a large number of free parameters to be estimated. This makes successful estimation dependent upon prior constraints that harness some natural properties of neuronal dynamics (e.g., neuronal activity cannot diverge exponentially to infinite values). A natural way to embody the requisite constraints is within a Bayesian framework. Dynamic causal models are therefore estimated using Bayesian estimators as described in Friston (2002). The estimation procedure provides the ''posterior density,'' that is, the probability distribution of a connectivity parameter in terms of its mean and standard deviation. For a given posterior density, the probability that an estimated parameter exceeds some specified threshold can then be computed. Unlike structural equation modeling, there are no limits on the number of connections that can be modeled because the assumptions and estimation procedures used by DCM are completely different, relying upon known inputs.

To summarize, DCM can be distinguished from extant approaches to brain connectivity in that it (a) frames the estimation problem in terms of ''designed perturbations,'' (b) makes inferences based on inferred neuronal states rather than the measured BOLD signal, (c) accommodates the nonlinear and dynamic aspects of neuronal interactions, and (d) uses a Bayesian framework that places no limit on the number of connections that can be modeled. DCM therefore represents a fundamental departure from conventional approaches to causal modeling in neuroimaging and brings the analysis of brain connectivity much closer to the analysis of regionally specific effects usually applied to neuroimaging data. For the operational details of DCM, see Appendix and Friston, Harrison, and Penny (2003).

Combining Functional Magnetic Resonance Imaging and Dynamic Causal Modeling

In the present study, we applied DCM to the fMRI data originally reported by Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000). Specifically, we investigated whether category effects in the occipital and temporal cortex are mediated by inputs from early visual cortex or superior/inferior parietal area. Whereas Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000) identified category effects in both hemispheres, we focused on the left hemisphere for computational expediency.

First, we performed a conventional Statistical Parametric Mapping (SPM) analysis independently for each

Table 1. Regions that Showed Greater Activation for Object Stimuli Relative to Scrambled Pictures (Category-Unresponsive) and Regions that Showed Greater Activation for Object Stimuli Relative to Scrambled Pictures and for One Object Category Relative to the Others (Category-Responsive)

	Subject 1	*Subject 2*	*Subject 3*	*Subject 4*	*Subject 5*
Category: Unresponsive					
Houses, Faces, and Chairs > Scrambles					
V3	−42,−88,−10 (**8.5**)	−44,−84,−10 (**8.8**)	−38,−80,−18 (**6.8**)	−40,−86,−8 (**8.0**)	−46,−86,−6 (**7.7**)
Superior parietal	−30,−62,72 (**7.6**)	−28,−62,72 (**8.4**)	−36,−68,62 (**7.8**)		−32,−50,70 (**8.0**)
Inferior parietal				−34,−74,48 (**8.7**)	
Category: Responsive					
Chairs > Faces and Houses					
Middle occipital (posterior)	−58,−74,−10 (4.2)	−30,−82,−10 (3.3)	−50,−80,−8 (**5.0**)		
Superior occipital			−54,−88,14 (**5.2**)	−35,−98,14 (3.6)	−60,−82,10 (**7.4**)
Faces > Houses and Chairs					
Middle occipital (posterior				−50,−92,−6 (**7.7**)	
Middle occipital (anterior)	−42,−64,−6 (**5.2**)				−46,−86,8 (3.6)
Inferior temporal		−58,−58,−18 (**6.0**)	−48,−66,−22 (**5.5**)		
Middle temporal		−64,−62,26 (**6.0**)			
Houses > Chairs and Faces					
Inferior fusiform	−24,−72,−12 (**7.5**)	−38,−62,−8 (**7.7**)	−34,−88,−4 (**5.5**)	−30,−80,−2 (4.7)	−26,−60,−12 (4.1)

Upper part: Regions that showed greater activation for object stimuli relative to scrambled pictures (i.e., chairs, faces, houses > scrambles) at $p <$ .05 (corrected for multiple comparisons)—Z scores are reported in parentheses. These regions did not show category effects, even when lowering the statistical threshold to $p < .05$ (uncorrected). Lower part: Regions that showed greater activation for object stimuli relative to scrambled pictures ($p < .05$ corrected for multiple comparisons) and for one object category relative to the others ($p < .001$ uncorrected)—Z scores for the category effects are reported in parentheses. All regions were identified, independently for each subject, using SPM analysis. Z scores that survived correction for multiple comparisons ($p < .05$) are reported in **bold**.

subject. This analysis identified regions that showed differential activity for object stimuli (i.e., houses, chairs, and faces) relative to scrambled pictures and for one object category relative to the others. V3 and the superior/inferior parietal area showed greater activation for object stimuli relative to scrambled pictures, but did not show any category effect (in Subjects 1–5; see upper part of Table 1). In contrast, a number of regions in the occipital and temporal gyri showed greater activation for object stimuli relative to scrambled pictures and differential activation for houses, faces, and chairs (in Subjects 1–5; see lower part of Table 1). The category effects reported by the Ishai et al. at group level were therefore replicated at an individual subject level when using SPM.

Second, we constructed a series of subject-specific dynamic causal models that comprised the house-, face-, and chair-responsive regions in the occipital and temporal cortex. V3 and the superior/inferior parietal area were also included as they expressed greater activation for houses, faces, and chairs relative to scrambled pictures but did not show any category effect. As represented in Figure 1, the dynamic causal model comprised intrinsic connections from V3 to the parietal cortex, from V3 to the category-responsive areas, and from the parietal cortex to the category-responsive areas. Bilinear terms were also specified to look at the influence of object category on the intrinsic connections from V3 to the category-responsive regions and from the parietal cortex to the category-responsive regions.

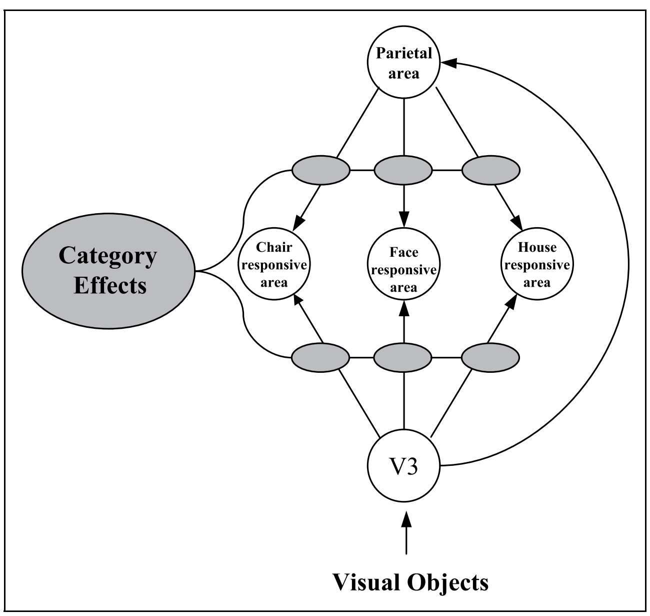

Figure 1. The dynamic causal model—which included V3, a superior/inferior parietal area, and the category-responsive regions in the occipital and temporal cortex. The vector ‘‘visual objects’’ encoded the presentation of visual objects (i.e., houses, faces, and chairs) and entered the model through the ‘‘input area’’ V3.

The stimulus function, which encoded the presentation of visual objects (i.e., houses, faces, and chairs), was entered into the dynamic causal model through the sensory area V3. The resulting perturbation was then allowed to propagate throughout the model via anatomical interconnections between V3 and the remaining regions.

DCM analysis estimated the intrinsic connections specified by the model and how these were influenced by the presentation of houses, faces, and chairs. The influence of object category on the intrinsic connections, modeled by the bilinear terms, was the primary focus of our DCM study. We hypothesized a significant influence of object category on the intrinsic connections that would account for the category effects observed in the occipital and temporal cortex. One possibility was that this influence would be expressed through the connections from V3 to the category-responsive areas—which would suggest bottom–up modulation. Another possibility was that the influence of object category on the connectivity parameters was expressed in the connections from parietal cortex to the category-responsive areas—thereby indicating top–down modulation. Finally, it was possible that object-specificity was conferred by connections from both V3 and parietal cortex.

RESULTS

Intrinsic Connections

The intrinsic connections from V3 to the category- responsive regions and the superior/inferior parietal area were significantly greater than 0 in all five subjects. In addition, the intrinsic connections from the superior/inferior parietal area to some of the category-responsive regions were significant in Subjects 3, 4, and 5. Results are shown graphically in Figure 2; intrinsic connections significant at 95% confidence are represented using solid lines and their estimates are reported in bold.

Influence of Object Category on the Intrinsic Connections

In all five subjects, category effects in the occipital and temporal cortex were mediated by inputs from V3. In other words, the intrinsic connectivity from V3 to the chair-responsive regions was stronger during the presentation of chairs than any other category; the intrinsic connectivity from V3 to the face-responsive regions was stronger during the presentation of faces than any other category; and the intrinsic connectivity from V3 to the house-responsive regions was stronger during the presentation of houses than any other category. In contrast, no subject showed an effect of object category on the intrinsic connections from the superior/inferior parietal area to the category-responsive regions. Results are shown graphically in Figure 2; bilinear terms significant

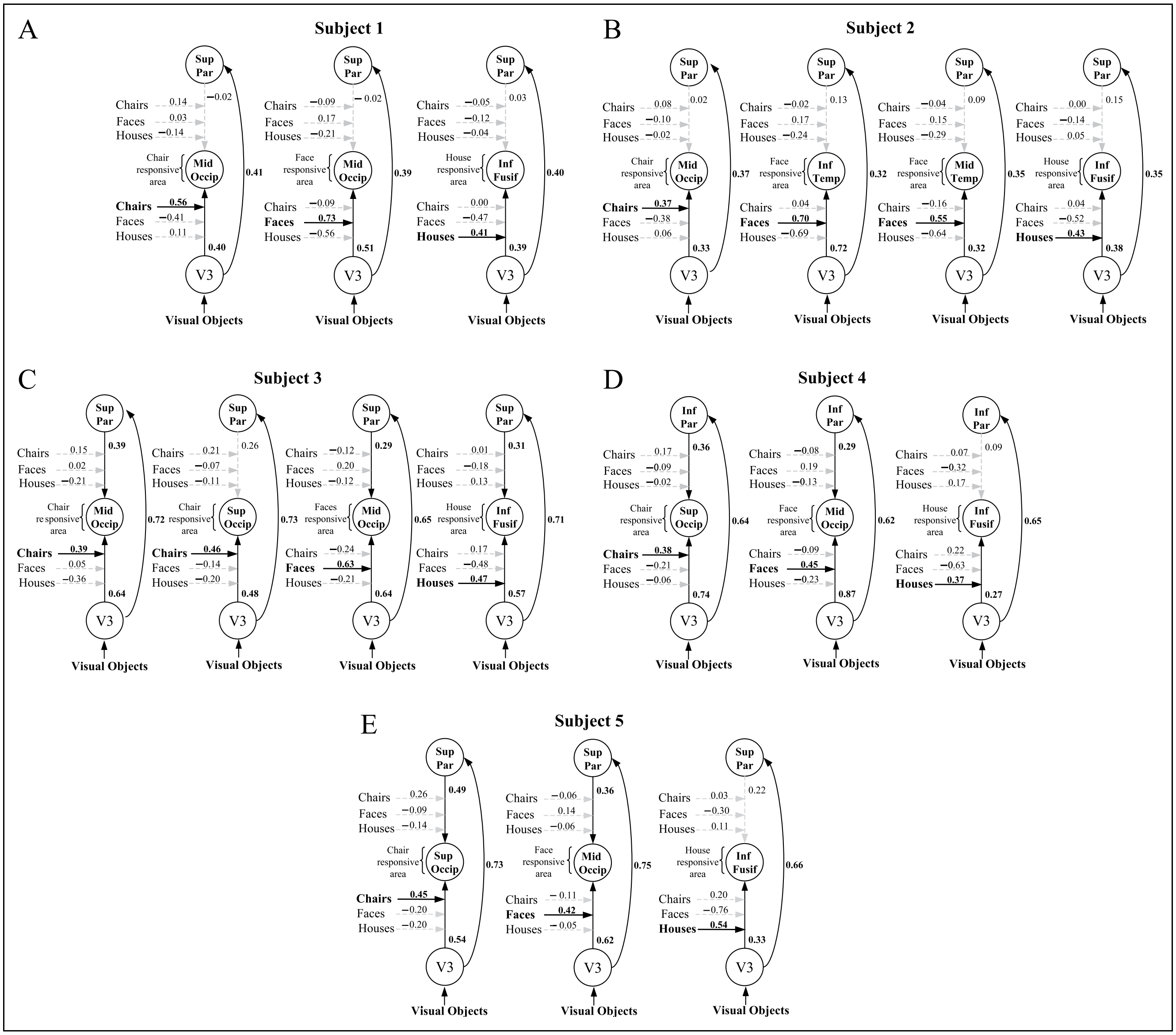

Figure 2. Results of the DCM analysis for each subject. Intrinsic connections and bilinear terms significant at 95% confidence are represented using solid lines and their estimates are reported in bold. It can be seen that, consistently across subjects, the intrinsic connectivity from V3 to the house-, face-, and chair-responsive regions was stronger during the presentation of houses, faces, and chairs, respectively. In contrast, the intrinsic connectivity from the superior/inferior parietal area to the category responsive regions appeared to be unaffected by the presentation of houses, faces, and chairs.

at 95% confidence are represented using solid lines and their estimates are reported in bold.

DISCUSSION

The present investigation represents the first attempt to explore the neuronal interactions that mediate category effects in the left occipital and temporal cortex. The study was motivated by the idea that functional specialization is not an intrinsic property of any region, but depends on both forward and backward connections (Friston & Price, 2001; McIntosh, 2000). Using the newly developed analytical technique DCM, we showed that the category effects reported by Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000) are associated with greater connectivity from early visual cortex. This finding indicates that category effects in the occipital and temporal cortex can be mediated by bottom–up mechanisms. This had been hypothesized in a number of earlier studies but had proved difficult to demonstrate using conventional analyses of regionally specific effects. In contrast, we found no evidence that the category effects reported by Ishai et al. are associated with greater connectivity from the parietal cortex. However, interpretation of this finding should be qualified, as we discuss below.

It may be interesting to consider the implications of our findings, together with the results reported by Ishai et al., for theories of category specificity. First, the representation of houses, faces, and chairs was not restricted to regions of cortex that responded exclusively to one category, rather activation was highly distributed across the occipital and temporal cortex. This finding is inconsistent with the hypothesis of segregated regions for different category-specific modules (see also Haxby et al., 2000). Second, within this distributed network, there were regions that responded preferentially to specific object categories, such as houses, faces, and chairs. This finding appears to be inconsistent with the hypothesis of an undifferentiated neural system for processing different object categories. Third, category effects in the occipital and temporal cortex were mediated by bottom–up mechanisms. This finding is inconsistent with the idea that category effects can be fully explained in terms of top–down attentional mechanisms. Rather, it appears to suggest that pictures of houses, faces, and chairs are associated with different visual features—which in turn may lead to functional specialization in the occipital and temporal cortex. However, this does not imply that category-related responses in the occipital and temporal cortex are passively driven by the visual features of the stimuli (see below).

In our investigation, we also found no evidence that the category effects reported by Ishai et al. are dependent on inputs from the parietal cortex. Interpretation of this finding should be cautious however, as we used a relatively simple dynamic causal model that comprised V3, superior/inferior parietal, and category-responsive regions only. The neural network that mediates category effects in the occipital and temporal cortex is likely to comprise a number of anatomical regions that were not included in our simple model. Primary candidates are the left prefrontal areas, which may be involved in semantic retrieval (Noppeney & Price, 2002), and medial anterior temporal areas, which may contribute to the integration of simple semantic features into a single object representation (Devlin, Moore, et al., 2002). These regions were not included in our dynamic causal model because Ishai et al. only acquired functional data from the back of the brain (see Methods). It is possible, however, that the prefrontal or the anterior temporal areas were coupled with the category-responsive areas in the occipital and temporal cortex, and that this coupling was influenced by object category. A critica consequence of these considerations is that, even if we found no evidence that category effects are mediated by inputs from the superior/inferior parietal area, we cannot discard the possibility that top–down mechanisms are involved in the category effects reported by Ishai et al. In fact, top–down modulation has been suggested by functional imaging data demonstrating category-related responses during word naming (Chao et al., 1999) and visual imagery (Ishai, Ungerleider, & Haxby, 2000). Furthermore, top–down modulation in category-responsive regions is supported by a task-dependent double dissociation between animals and tools (Devlin, Moore, et al., 2002).

In summary, it is likely that the bottom–up modulation identified by our DCM analysis is not the only way in which category effects in the occipital and temporal cortex can be mediated. We suggest that category effects, like those reported by Ishai et al., can be mediated either by bottom–up or top–down mechanisms—depending on the context. For instance, passive viewing and visual imagery may well identify the same category effects but these could be mediated by bottom–up and top–down mechanisms, respectively. This hypothesis can be tested empirically by applying our DCM analysis to data sets obtained using a range of experimental paradigms that do and do not rely on pictorial stimuli. In other words, an experimental paradigm should be used that involves explicit manipulation of top–down mechanisms. An alternative possibility is that the bottom–up modulation identified by our DCM analysis was, in turn, mediated by inputs from frontal or anterior temporal regions to early visual areas. This hypothesis can also be tested empirically with DCM by using a data set that includes functional information from these regions.

Finally, the present study illustrates the use of DCM to estimate the influence that one region exerts over another and how this is affected by experimental manipulation. By using a Bayesian framework, that places no limits on the number of connections that can be modeled, we were able to test models that would be impossible to characterize using existing methods based on multiple regression. Furthermore, by treating the measured hemodynamic responses as evoked by known experimental inputs, we avoided the assumption that the measured hemodynamic responses are driven by endogenous or intrinsic noise. Given that the vast majority of functional imaging studies rely on designed experiments, we consider DCM a potentially useful complement to existing techniques.

METHODS

Experimental Design

Six healthy right-handed subjects performed passive viewing and delayed match-to-sample tasks. In the passive viewing task, single stimuli (gray-scale photographs of houses, faces, and chairs) were presented at a rate of two per second. Scrambled pictures of houses, faces, and chairs were also presented with the same temporal sequence as a control. In the delayed matching task, a single-sample stimulus (presented for 1.5 sec) was followed by a pair of choice stimuli (presented for 2 sec). Subjects indicated which choice stimulus matched the sample stimulus by pressing a button with the right or left thumb. In the control task, scrambled pictures of houses, faces, and chairs were presented with the same temporal sequence. Here subjects responded to the presentation of each pair of scrambled items by pressing both right and left buttons simultaneously. See Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000) for details.

Data Acquisition

A 1.5-T General Electric Signa scanner was used to acquire blood oxygen level-dependent T2*-weighted MRI signal (TR = 3 sec) and high-resolution full-volume structural images. Each functional image comprised 18 contiguous, 5-mm thick coronal slices to cover occipital, parietal, and posterior temporal cortex. See Ishai et al. (1999) and Ishai, Ungerleider, Martin, et al. (2000) for details.

Data Analysis

Statistical Parametric Mapping

SPM analysis (Friston, Holmes, et al., 1995) was performed using SPM2 software (Wellcome Department of Imaging Neuroscience, London, UK; http://www.fil.ion.ucl.ac.uk), running under Matlab 6 (Mathworks, Sherbon, MA). Functional images were realigned with an iterative method and coregistered to the individual T1-weighted structural images. The T1-weighted structural images were normalized in the space of Talairach and Tournoux (1988) using nonlinear basis functions (Friston, Ashburner, et al., 1995) and the resulting "warp" parameters were then applied to the functional images. These were spatially smoothed with a Gaussian filter of 3.75. After preprocessing, a series of subject-specific models were created to characterize the hemodynamic response under each experimental condition. The data were high-pass filtered using a set of discrete cosine basis functions with a cutoff period of 128 sec. The analysis used the general linear model to identify regions that showed differential activity (a) for object stimuli (i.e., houses, chairs, and faces) relative to scrambled pictures and (b) for houses, chairs, and faces (e.g., houses > chairs and faces, etc.). To maximize sensitivity to category effects in the occipital and temporal cortex, results are reported at $p < .001$ (uncorrected for multiple comparisons) with an extent threshold for each cluster of 5 voxels.

Dynamic Causal Modeling

DCM (Friston et al., in press) was also performed using SPM2 software running under Matlab 6. DCM was performed on subjects 1–5 only, as SPM did not detect any significant effects in subject 6. A series of subject-specific dynamic causal models was constructed that comprised V3, a superior/inferior parietal area, and category-responsive regions in the left hemisphere. Regions (8 mm radius) were selected independently for each subject using maxima of the SPM$\{T\}$ obtained using the conventional SPM analysis. Principal eigenvariates were extracted from all regions and entered into the DCM analysis to estimate intrinsic connections and how these were influenced by the presentation of houses, faces, and chairs. Bayesian inferences were based upon the probability that the coupling parameters exceeded 0. Inferences were made at 95% confidence. It should be noted that correction for multiple comparisons is not required in DCM as there are no null hypotheses tested in a classical sense. Rather, the probability that an estimated connectivity parameter lies in a certain range of values (e.g., $0 \rightarrow \infty$) is computed.

APPENDIX

Here we present the operational details upon which DCM rests (also see Friston et al., 2003). In brief, DCM is a fairly standard nonlinear system identification procedure using Bayesian estimation of the parameters of deterministic input–state–output dynamic systems. The estimation conforms to the posterior density analysis under Gaussian assumptions described in Friston (2002). This posterior density analysis finds the most likely coupling parameters given the data by performing a gradient assent on the log posterior. The log posterior requires both likelihood and prior terms. The likelihood obtains from Gaussian assumptions about the errors in the observation model implied by the DCM. This likelihood or forward model is described in the next subsection. The priors on the coupling and hemodynamic parameters obtain using a fully Bayesian approach as described in the second subsection. In the third subsection, we show that, by combining the likelihood with the priors, one can form an expression for the posterior density that is used in the estimation. Finally, we consider the relationship between DCM and conventional analyses.

Dynamic Causal Models

A dynamic causal model is a multiple-input–multiple-output system that comprises m inputs and l outputs with one output per region. The m inputs correspond to designed causes (e.g., a stick stimulus function) and are the same as those used to form design matrices in conventional analyses of fMRI data. The l outputs correspond to the observed BOLD signal and would normally be taken as the average or first eigenvariate of key regions, selected on the basis of a conventional analysis. Each region in a dynamic causal model has five state variables. Central to the estimation of effective connectivity or coupling parameters is the average neuronal activity. This state variable is a function of the neuronal states of other brain regions. The remaining state variables are of secondary importance and correspond to the biophysical states of the hemodynamic model presented in Mechelli et al. (2001) and Friston et al. (2000). These biophysical states are required to compute the

BOLD response from the average neuronal activity and are not influenced by the neuronal states of other brain regions. We will deal first with the equations for the neuronal state variable and then briefly reprise the hemodynamic model for each region.

Neuronal State Equations

Restricting ourselves to the neuronal states $z = [z_1, \ldots, z_l]^T$ one can posit any arbitrary form or model for effective connectivity:

$$\dot{z} = F(z, u, \theta) \tag{1}$$

where F is some nonlinear function describing the neurophysiological influences that activity in all l brain regions z and inputs u exert upon changes in the others. θ are the parameters of the model whose posterior density (i.e., the probability distribution in terms of its mean and standard deviation) we require for inference. The bilinear approximation of equation 1 provides a natural and useful reparameterization in terms of effective connectivity:

$$\begin{aligned}
\dot{z} &\approx Az + \sum u_j B^j z + Cu \\
&= \left(A + \sum u_j B^j\right) z + Cu \\
A &= \frac{\partial F}{\partial z} = \frac{\partial \dot{z}}{\partial z} \\
B^j &= \frac{\partial^2 F}{\partial z \partial u_j} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z} \\
C &= \frac{\partial F}{\partial u}
\end{aligned} \tag{2}$$

The Jacobian or connectivity matrix A represents the first-order connectivity among the regions in the absence of input. In DCM, a response is defined in terms of a change in activity with time $\dot{z}$. Effective connectivity is the influence that one neuronal system exerts over another in terms of inducing a response $\partial\dot{z}/\partial z$. This first-order connectivity can be thought of as the intrinsic coupling in the absence of experimental perturbations. This state depends on the experimental design and therefore the intrinsic coupling is specific to each experiment. Matrix B^j is effectively the change in intrinsic coupling induced by the jth input. They encode the input-sensitive changes in $\partial\dot{z}/\partial z$ or, equivalently, the modulation of effective connectivity by experimental manipulations. Because B^j is a second-order derivative, this term is referred as bilinear. Finally, the matrix C embodies the extrinsic influences of inputs on neuronal activity. The parameters $\theta^c = \{A, B^j, C\}$ are the connectivity or coupling matrices that we wish to identify and define the functional architecture and interactions among brain regions at a neuronal level. In DCM, the units of connections are per unit time and therefore correspond to rates. Because we are in a dynamical setting, a strong connection means an influence that is expressed quickly or with a small time constant. The neuronal activity in each region causes changes in volume and deoxyhemoglobin to engender the observed BOLD response y as described next.

Hemodynamic Model

The remaining state variables of each region are biophysical states engendering the BOLD signal and mediate the translation of neuronal activity into hemodynamic responses. These biophysical states $\{s, f, v, q\}$ comprise a vasodilatory signal, normalized flow, normalized venous volume, and normalized deoxyhemoglobin content. These state variables are a function of, and only of, the neuronal state of each region. The equations have been described previously (Mechelli et al., 2001; Friston et al., 2000) and constitute a hemodynamic model that embeds the Balloon–Windkessel component (Mandeville et al., 1999; Buxton, Wong, & Frank, 1998). Additional biophysical parameters $\theta^h = \{\kappa, \gamma, \tau, \alpha, \rho\}$ in the hemodynamic model comprise rate of signal decay, rate of flow-dependent elimination, hemodynamic transit time, Grubb's exponent (Grubb, Rachael, Euchring, & Ter-Pogossian, 1974), and resting oxygen extraction fraction.

Combining the neuronal states with the biophysical states gives us a full forward model

$$\begin{aligned}
\dot{x} &= f(x, u, \theta) \\
y &= \lambda(x)
\end{aligned} \tag{3}$$

with state variables $x = \{z, s, f, v, q\}$ and parameters $\theta = \{\theta^c, \theta^h\}$. For any set of parameters and inputs, the state equation can be integrated and passed through the output nonlinearity to give the predicted response $h(u, \theta)$. This integration can be made quite expedient by capitalizing on the sparsity of stimulus functions commonly employed in fMRI designs (see Friston, 2002).

The forward model can be made into an observation model by adding error and confounding or nuisance effects $X(t)$ to give $y = h(u, \theta) + X\beta = \varepsilon$. Here β is the unknown coefficient of the confounds. Following the approach described in Friston (2002), we note

$$\begin{aligned}
y - h(u, \eta_{\theta|y}) &\approx J\Delta\theta + X\beta + \varepsilon \\
&= [J, X]\begin{bmatrix}\Delta\theta \\ \beta\end{bmatrix} + \varepsilon \\
\Delta\theta &= \theta - \eta_{\theta|y} \\
J &= \frac{\partial h(u, \eta_{\theta|y})}{\partial\theta}
\end{aligned} \tag{4}$$

This local linear approximation then enters an EM scheme as described previously

Until convergence {

E-step

$$
\begin{aligned}
J &= \frac{\partial h(\eta_{\theta|y})}{\partial\theta} \\
\bar{y} &= \begin{bmatrix} y - h(\eta_{\theta|y}) \\ \eta_\theta - \eta_{\theta|y} \end{bmatrix}, \bar{J} = \begin{bmatrix} J & X \\ 1 & 0 \end{bmatrix}, \bar{C}_\varepsilon = \begin{bmatrix} \sum \lambda_i Q_i & 0 \\ 0 & C_\theta \end{bmatrix} \\
C_{\theta|y} &= \left(\bar{J}^T \bar{C}_\varepsilon^{-1} \bar{J}\right)^{-1} \\
\begin{bmatrix} \Delta\eta_{\theta|y} \\ \eta_{\beta|y} \end{bmatrix} &= C_{\theta|y}\left(\bar{J}^T \bar{C}_\varepsilon^{-1} \bar{y}\right) \\
\eta_{\theta|y} &\leftarrow \eta_{\theta|y} + \Delta\eta_{\theta|y}
\end{aligned} \tag{5}
$$

M-step

$$
\begin{aligned}
P &= \bar{C}_\varepsilon^{-1} - \bar{C}_\varepsilon^{-1}\bar{J}C_{\theta|y}\bar{J}^T\bar{C}_\varepsilon^{-1} \\
\frac{\partial F}{\partial \lambda_i} &= -\tfrac{1}{2}tr\{PQ_i\} + \tfrac{1}{2}\bar{y}^T P^T Q_i P \bar{y} \\
\left\langle \frac{\partial^2 F}{\partial \lambda_{ij}^2} \right\rangle &= -\tfrac{1}{2}tr\{PQ_iPQ_j\} \\
\lambda &\leftarrow \lambda - \left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle^{-1} \frac{\partial F}{\partial \lambda}
\end{aligned}
$$

The **M** step updates error covariance parameters λ. The prediction and observations encompass the entire experiment. They are therefore large $ln \times 1$ vectors whose elements run over regions and time. Although the response variable could be viewed as a multivariate time series, it is treated as a single observation vector, whose error covariance embodies both temporal and interregional correlations. $C_\varepsilon = V \otimes \sum(\lambda) = \sum \lambda_i Q_i$. This covariance is parameterized by some covariance parameters λ. These correspond to region-specific error variances assuming the same temporal correlations $Q_i = V \otimes \sum_i$ in which $\sum_i$ is a $l \times l$ sparse matrix with the ith leading diagonal element equal to 1.

Equation 5 enables us the estimate the conditional moments of the coupling parameters (and the hemodynamics parameters) plus the parameters controlling observation error. However, to proceed we need to specify the priors.

Priors

Here we use a fully Bayesian approach because (a) there are clear and necessary constraints on neuronal dynamics that can be used to motivate priors on the coupling parameters and (b) empirically determined priors on the biophysical hemodynamic parameters are relatively easy to specify. We will deal first with priors on the coupling parameters.

Priors on the Coupling Parameters

It is self-evident that neuronal activity cannot diverge exponentially to infinite values. Therefore, we know that, in the absence of input, the dynamics must to return to a stable mode. This means the largest real component of the eigenvalues of the intrinsic coupling matrix cannot exceed zero. We use this constraint to establish a prior density on the coupling parameters A that ensures the system is dissipative.

If the largest real eigenvalue (Lyapunov exponent) is less than zero, the stable mode is a point attractor. If the largest Lyapunov exponent is zero, the system will converge to a periodic attractor with oscillatory dynamics. Therefore, it is sufficient to establish a probabilistic upper bound on the interregional coupling strengths; imposed by Gaussian priors that ensures the largest Lyapunov exponent is unlikely to exceed zero. If the prior densities of each connection are independent, then the prior density can be specified in terms of a variance for the off-diagonal elements of A. This variance can then be chosen to render the probability of the principal exponent exceeding zero, less than some suitably small value (see Friston et al., 2003, for details).

Hemodynamic Priors

The hemodynamic priors are based on those used in Friston (2002). In brief, the mean and variance of posterior estimates of the five biophysical parameters were computed over 128 voxels using the single-word presentation data presented in the next section. These means and variances were used to specify Gaussian priors on the hemodynamic parameters.

Combining the prior densities on the coupling and hemodynamic parameters allows us to express the prior probability of the parameters in terms of their prior expectation η_θ and covariance C_θ. Having specified the priors, we are now in a position to form the posterior and proceed with estimation using equation 5.

Inference

As noted above, the estimation scheme is a posterior density analysis under Gaussian assumptions (see Friston, 2002, for details). In short, the estimation scheme provides the approximating Gaussian posterior density of the parameters $q(\theta)$ in terms of its expectation $\eta_{\theta|y}$ and covariance $C_{\theta|y}$. The expectation is also known as the posterior mode or maximum a posteriori (MAP) estimator. The marginal posterior probabilities are then used for inference that any particular parameter or contrast of parameters $c^{T}\eta_{\theta|y}$ (e.g., average) exceeded a specified threshold γ.

$$p = \phi_N \left(\frac{c^T \eta_{\theta|y} - \gamma}{\sqrt{c^T C_{\theta|y} c}} \right) \tag{6}$$

ϕ_N is the cumulative normal distribution. The units of these parameters are hertz or per second (or adimensional if normalized) and the thresholds are specified as such. In dynamical modeling, strength corresponds to a fast response with a small time constant.

Relationship to Conventional Analyses

Conventional analyses of fMRI data using linear convolution models are a special case of dynamic causal models using a bilinear approximation. This is important because it provides a direct connection between DCM and classical models. If we allow inputs to be connected to all regions and discount interactions among regions by setting the prior variances on A and B to zero we produce a set of disconnected brain regions or voxels that respond to and only to extrinsic input. The free parameters of interest reduce to the values of C, which reflect the ability of input to excite neural activity in each voxel. By further setting the prior variances on the self-connections (i.e., scaling parameter) and those on the hemodynamic parameters to zero, we end up with a single-input–single-output model at each and every brain region that can be reformulated as a convolution model as described in Friston (2002). The key point here is that the general linear models used in typical data analyses are special cases of bilinear models that embody more assumptions. These assumptions enter through the use of highly precise priors that discount interactions among regions and prevent any variation in biophysical responses.

Acknowledgments

A. Mechelli is supported by MH64445 from the National Institutes of Health (USA). C. J. Price, U. Noppeney, and K. J. Friston are supported by The Wellcome Trust.

Reprint requests should be sent to Andrea Mechelli, Wellcome Department of Imaging Neuroscience, Institute of Neurology, 12 Queen Square, London WC1N 3BG, UK, or via e-mail: andream@fil.ion.ucl.ac.uk.

REFERENCES

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The Balloon model. *MRM, 39,* 855–864.

Cappa, S. F., Perani, D., Schnur, T., Tettamanti, M., & Fazio, F. (1998). The effects of semantic category and knowledge type on lexical–semantic access: A PET study. *Neuroimage, 8,* 350–359.

Caramazza, A., & Shelton, J. (1998). Domain-specific knowledge systems in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.

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

Devlin, J. T., Moore, C. J., Mummery, C. J., Gorno-Tempini, M. L., Phillips, J. A., Noppeney, U., Frackowiak, R. S. J., Friston, K. J., & Price, C. J. (2002). Anatomic constraints on cognitive theories of category specificity. *Neuroimage, 15,* 675–685.

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, J., & Tyler, L. K. (2002). Is there an anatomical basis for category specificity? Semantic memory studies in PET and fMRI. *Neuropsychologia, 40,* 54–75.

Friston, K. J. (2002). Bayesian estimation of dynamical systems: An application to fMRI. *Neuroimage, 16,* 513–530.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B, Heather, J. D., & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Human Brain Mapping, 2,* 1–25.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage, 19,* 1273–1302.

Friston, K. J., Holmes, A., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging. A general linear approach. *Human Brain Mapping, 2,* 189–210.

Friston, K. J., & Price, C. J. (2001). Dynamic representations and generative models of brain function. *Brain Research Bulletin, 54,* 275–285.

Grubb, R. L., Rachael, M. E., Euchring, J. O., & Ter-Pogossian, M. M. (1974). The effects of changes in PCO2 on cerebral blood volume, blood flow and vascular mean transit time. *Stroke, 5,* 630–639.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.

Haxby, J. V., Ishai, A., Chao, L. L., Ungerleider, G., & Martin, A. (2000). Object–form topology in the ventral temporal lobe. *Trends in Cognitive Sciences, 4,* 3–4.

Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron, 28,* 979–990.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, U.S.A., 96,* 9379–9384.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience, 12,* 35–51.

Mandeville, J. B., Marota, J. J., Ayata, C., Zararchuk, G., Moskowitz, M. A., Rosen, B., & Weisskoff, R. M. (1999). Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *Journal of Cerebral Blood Flow and Metabolism, 19,* 679–689.

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural Networks, 13,* 861–870.

Mechelli, A., Price, C. J., & Friston, K. J. (2001). Nonlinear coupling between evoked rCBF and BOLD signals: A simulation study of hemodynamic responses. *Neuroimage, 14,* 862–872.

Noppeney, U., & Price, C. J. (2002). A PET study of stimulus- and task-induced semantic processing. *Neuroimage, 15,* 927–935.

Talairach, J., & Tournoux, P. (1988). *A co-planar stereotactic atlas of the human brain.* Stuttgart: Thieme.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75,* 195–231.

Warrington, E., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829–853.