PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

On imputing function to structure from the behavioural effects of brain lesions

Malcolm P. Young, Claus C. Hilgetag and Jack W. Scannell

Phil. Trans. R. Soc. Lond. B 2000 **355**, 147-161
doi: 10.1098/rstb.2000.0555

THE ROYAL SOCIETY

On imputing function to structure from the behavioural effects of brain lesions

Malcolm P. Young*, Claus-C. Hilgetag and Jack W. Scannell

Neural Systems Group, Department of Psychology, University of Newcastle upon Tyne, Ridley Building,
Newcastle upon Tyne NE1 7RU, UK

What is the link, if any, between the patterns of connections in the brain and the behavioural effects of localized brain lesions? We explored this question in four related ways. First, we investigated the distribution of activity decrements that followed simulated damage to elements of the thalamocortical network, using integrative mechanisms that have recently been used to successfully relate connection data to information on the spread of activation, and to account simultaneously for a variety of lesion effects. Second, we examined the consequences of the patterns of decrement seen in the simulation for each type of inference that has been employed to impute function to structure on the basis of the effects of brain lesions. Every variety of conventional inference, including double dissociation, readily misattributed function to structure. Third, we tried to derive a more reliable framework of inference for imputing function to structure, by clarifying concepts of function, and exploring a more formal framework, in which knowledge of connectivity is necessary but insufficient, based on concepts capable of mathematical specification. Fourth, we applied this framework to inferences about function relating to a simple network that reproduces intact, lesioned and paradoxically restored orientating behaviour. Lesion effects could be used to recover detailed and reliable information on which structures contributed to particular functions in this simple network. Finally, we explored how the effects of brain lesions and this formal approach could be used in conjunction with information from multiple neuroscience methodologies to develop a practical and reliable approach to inferring the functional roles of brain structures.

Keywords: double dissociation; structure–function relationships; corticocortical connections; thalamocortical connections; inference; neuroinformatics

1. INTRODUCTION

t is a long-standing premise in brain science (e.g. 'lechsig 1905; Meynert 1890) that understanding how the rain is organized structurally will inform understanding f how it works. An important motivation behind much xperimental neuroanatomy, for example, has been the ntuition that structure–function relationships are of gnal importance in the brain, and that investigations of urely anatomical aspects of the brain could have a hysiological significance well beyond their actual subject atter. In many respects, this premise has been amply orne out, and the approach that derives from it has ucceeded spectacularly: very few neurophysiologists ould now find their work possible without the wide ariety of anatomically derived information that frames eir understanding of the systems they investigate. In ther respects, structure–function relationships at many cales of the nervous system have remained opaque and lusive. The well-known mismatch, for example, between ortical neurons' morphological extent and complexity nd the localized physiological properties that neuro-hysiologists report (Douglas & Martin 1991) has only ecently begun to give way (Douglas & Martin 1994; Douglas *et al*. 1996). Similarly, at the level of whole systems in the brain, the extent and complexity of cortico- and thalamocortical networks has been difficult to relate clearly to the functional properties of the network or of its constituent structures. This latter difficulty has also recently begun to give way, evidenced by the ability of analyses of these complex networks to predict successfully the location of cells with specific physiological properties (e.g. Scannell *et al*. 1996, 1997; cf. Merabet *et al*. 1998), to account for the distribution of particular kinds of selectivity by reference to the structure of part of the network (Burns & Young, this issue; Hilgetag *et al*. 1996; Hilgetag, Burns, O'Neill, Scannell & Young, this issue), and to account for the spatial distribution of activity across the areas of the cortex after localized experimental disinhibition (Kötter & Sommer, this issue; Stephan, Hilgetag, Burns, O'Neill, Young & Kötter, this issue).

These explicit systems-level structure–function relationships reveal parts of a causal bridge between connectional anatomy and physiological function. However, they do not yet directly inform the structure–function relationships that have been of most interest to behavioural neuroscientists. One object of that discipline is to try to identify the specific behavioural or cognitive functions mediated by specific anatomical structures by damaging

Author for correspondence (m.p.young@ncl.ac.uk).

he structures and observing the effects of this damage on ehaviour. Thus this aim is to impute specific function to pecific structures on the basis of the effects of lesions of hose structures. It is not unreasonable to think that pecific lesions have their effects on behaviour through heir effects on the network of connections in the brain, nd so on other brain structures. Yet the link between onnectivity and lesion effects remains almost completely paque.

We are interested in whether a mathematical and omputational bridge can be built between connectivity nd the behavioural effects of lesions in brain structures. uch a bridge could aid prediction, the reliability of nferences from lesion effects, and could begin to provide framework in which the multiple sources of information hat bear upon the function of a brain system, such as its onnectivity, neurophysiology, gross activation and the ffects of lesions of its structures, could inform one nother formally, and hence lead towards better understanding. We assume that one end of a bridge between onnectivity and the functional effects of lesions must be nchored on information about the connections between rain structures. Neuroinformatic studies of neuronatomical connectivity therefore formed our starting oint. We developed the link between connectivity and esion effects in the following ways, each of which is the ubject of one of the sections below.

First, recent demonstrations of structure–function elationships have employed simple integrative mechanisms to successfully relate connection data to information n the spread of activation (Kötter & Sommer, this ssue), and to account simultaneously for intact, lesioned nd several kinds of paradoxically restored orientating unction (Hilgetag, Burns, O'Neill, Scannell & Young, his issue). Together, these problems offer constraints from everal different experimental sources, suggesting that he integrative mechanisms that link them are a useful asis for initial modelling of the relationships between rain structures, including those perturbed by lesions. ccordingly, we began by selecting a system in which onnectivity has been well studied, the thalamocortical ystem of the cat (Scannell *et al.* 1999), and, using the ntegrative mechanisms that underlay the structure–unction relationships just described, investigated the istribution of activity decrements that followed imulated damage to elements of the thalamocortical etwork. Second, we examined the consequences of the atterns of decrement seen in the simulation for each type f inference that has been employed to impute function to tructure on the basis of the effects of brain lesions. Third, e tried to derive a more reliable framework of inference or imputing function to structure, by clarifying concepts f structure and function, and deriving a more formal ramework based on concepts capable of mathematical pecification. Fourth, we applied this framework to nferences about function relating to a simple network hat reproduces intact, lesioned and paradoxically estored orientating behaviour (Hilgetag, Burns, O'Neill, cannell & Young, this issue), and show that lesion effects an in some circumstances be used to recover reliable nformation on which brain structures contribute to particular behavioural functions. Finally, we explore how a eliable approach to inferring the functional role of brain structures from the effects of lesions to them might be further developed.

2. MODELLING DAMAGE IN A COMPLEX NETWORK

To explore the general effects of lesions on a complex network of cortical areas and thalamic nuclei, we have made a number of simple models based on experimentally reported thalamo-corticocortical connectivity. The connection data that we used, which include the extrinsic connections linking nearly all the areas of the cerebral cortex and nuclei of the thalamus (figure 1*a*), were collated by Scannell *et al.* (1999) and are available at (www.flash.ncl.ac.uk/ptrs/cat_cor_thal.htm). The integrative mechanisms used to model the dynamics of activity in individual stations and the propagation of activity through the network were inspired by, and closely related to, the mechanisms used successfully elsewhere to link empirically reported connectivity to the empirically reported propagation of activity (Kötter & Sommer, this issue), and connectivity and orientating behaviour (Hilgetag, Burns, O'Neill, Scannell & Young, this issue).

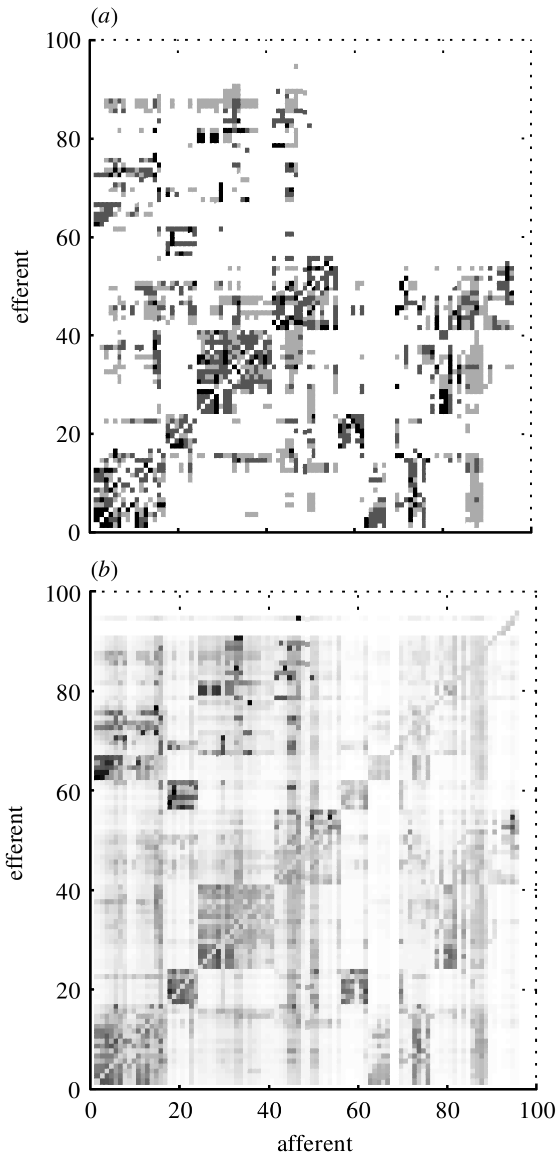
The present report concerns only the simplest model we have constructed. In the model, the mean level of activity in each cortical area or thalamic nucleus was represented as the level of activation of a unit. The pattern of connections between the units was derived from the known pattern of extrinsic connections between cortical areas and thalamic nuclei, so that each unit represented a particular cortical area or thalamic nucleus. The input to each unit, x_i, was given by equation (1), where $W_{j,i}$ was the connection weight of the jth to the ith unit, z_j was the activation of the jth unit, and g_i was the gain of the ith unit.

$$x_i = g_i \sum_j (W_{j,i} \times z_j). \tag{1}$$

The activation of each unit simply depended on its instantaneous level of input, x_i. Activation was calculated using a sigmoidal activation function, and could range between 0 and 1 (equation (2)). Parameter a (the offset of the activation function) was set to 0.5 and parameter k (the slope of the activation function) set to 3. The gain of each unit, g_i, was adjusted so that activation, z_i, settled to an equilibrium state of 0.5.

$$z_i = \frac{1}{1 + \mathrm{e}^{k(a-x_i)}}. \tag{2}$$

As the levels of gain were adjusted for each unit, the model network approached a state of equilibrium. When equilibrium was achieved, the gain for each unit was fixed. The adjustable gain was simply a scaling procedure, so that areas with many inputs did not remain at much higher levels of activation than areas with few inputs. Very similar results were obtained with fixed levels of gain. We then made 'lesions' in the network, by removing each unit in turn. We recorded the level of activation in all the other units in the network following each lesion, and this is shown in figure 1*b*. The simulation was run in MATLAB.



igure 1. Direct connections and lesion effects in a simple ıodel of the thalamo-corticocortical network. (*a*) The weights f the direct connections between 55 units representing articular cortical areas (*x*- and *y*-axes, structures 1 to 55) nd 41 units representing particular thalamic nuclei ι- and *y*-axes, structures 56 to 96). White, light grey, dark rey and black squares represent connections with weights of , 1, 2 and 3, respectively. The weights agree with the rank rder of the densities of the corresponding anatomical rojections. Note that there are no direct neural projections etween the units representing thalamic nuclei, as thalamic uclei do not have direct neural connections with each other. ‿eading vertically up the matrix shows the weights of each reas' inputs. Reading horizontally across the matrix shows ıe weights of each area's outputs. (*b*) The impact on the etwork of a lesion in each unit. The colours in the matrix epresent the level of activity after the lesion divided by the ·vel of activity prior to the lesion. White squares indicate no hange, darker squares represent stronger suppression. ‿eading horizontally shows the sensitivity of each unit to ·sions elsewhere. Reading vertically shows the effect of a ·sion in a unit on other units in the network. There is a good orrelation between the impact of lesions and the pattern of irect projections between units. However, even in this highly ımplified model, lesions have influences on structures to ·hich they do not send direct projections. This is particularly

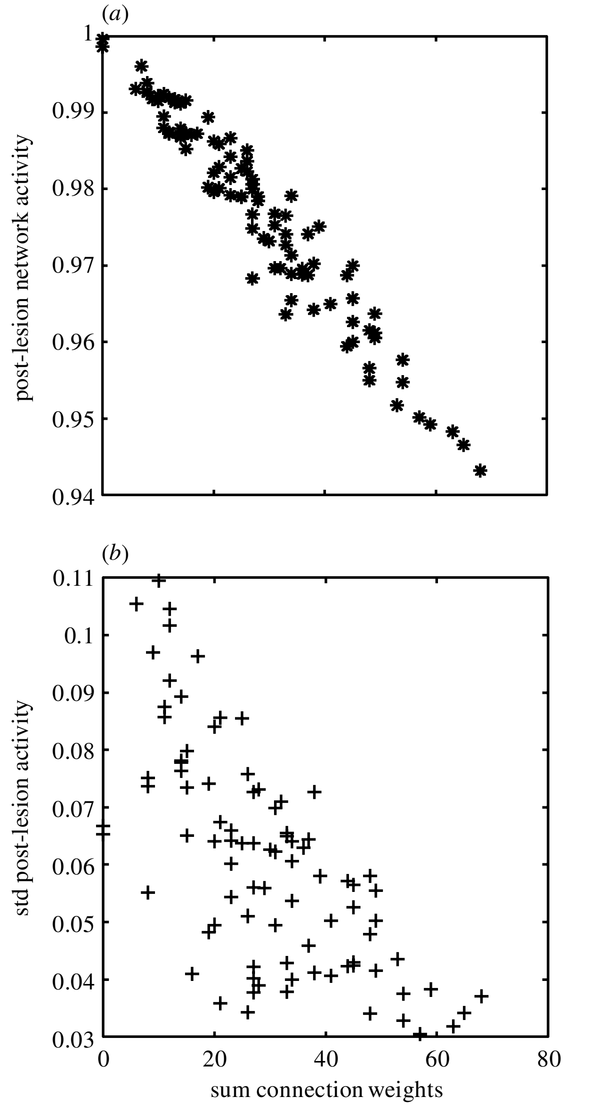
Figure 1*b* shows that lesions had an impact well beyond the unit that was lesioned. There was the expected high correlation between the distant impact of lesions and the pattern of direct projections between units. However, even with the simple integrative mechanisms employed in the model, the effects of lesions spread well beyond the units that received direct projections from the lesioned structure. This is most clearly demonstrated in the region of figure 1*b* that shows interactions between the thalamic units (units 55–96). By comparing this region of figures 1*a* and 1*b*, it is clear that the thalamic units influenced one another in the absence of any direct projections between them. Similar effects occurred between cortical units, but the profusion of direct corticocortical connections made this feature less obvious. Hence, even in a network with simple integrative mechanisms, lesions had effects that showed a complex dependence on the pattern of extrinsic connections between stations.

Figure 1*b* illustrates the spread of the indirect, 'network' effects of lesions. Figure 2 makes clear two other important interactions between connectivity and the effects of lesions, namely the different impacts of lesions in particular structures on activity in the network as a whole, and the different vulnerability exhibited by particular structures to lesions made elsewhere in the network. Figure 2*a* shows the ratio of pre- and post-lesion activity in the network following lesions in different units, against the number of connections possessed by the lesioned structure. It is clear that the number of connections that a unit had, expressed as the sum of its connection weights, strongly influenced the impact of a lesion in that unit on the activity in the rest of the network. Figure 2*b* shows how a unit's vulnerability to lesions elsewhere in the network also depended on the number and nature of connections that the unit made. Units that connect relatively widely tended to be suppressed by lesions in any of a large number of other structures, but the magnitude of the suppression was reasonably constant, no matter where the distant lesion was made. In contrast, units with relatively few connections had very variable vulnerability. They were very heavily suppressed by lesions in the few structures with which they were connected, but were much less sensitive to lesions in the many structures with which they did not connect. Hence, the number of connections possessed by a structure was an important determinant both of the impact that lesions of that structure had upon the network, and of the vulnerability of the structure to being affected by lesions made elsewhere. In the empirically derived thalamo-corticocortical network there is a high degree of variability in the number of extrinsic connections made by different cortical areas and thalamic nuclei (Scannell *et al*. 1999), suggesting that the impact of, and vulnerability of structures to, lesions will be highly variable between structures.

3. CONVENTIONAL INFERENCE

In §2, we examined the propagation of the effects of simulated lesions through the thalamocortical model and

Figure 1. (*Cont.*) evident in the region of the matrix from units 55 onward, where lesions in the 'thalamic' units influence other 'thalamic' units in the absence of any direct projections between them.



igure 2. Connectivity influences lesion impact and lesion ulnerability. (*a*) Relative activity in the network following :sions depends on the connectivity of the lesioned unit. The -axis shows total activity in network after a lesion divided by otal activity before the lesion. The *x*-axis shows the sum of ne connection weights of the lesioned unit. Lesions to units ith more connections have a larger impact on activity in the est of the network. (*b*) Vulnerability to lesions depends on ne units' connections. The *y*-axis provides a measure of the ariability in sensitivity to lesions: the standard deviation of ne activity in the unit following lesions elsewhere. The *x*-axis nows the sum of connection weights of the intact unit whose ctivity is measured. Units that make few connections have ery variable vulnerability. They escape the consequences of :sions in units to which they do not connect, but are severely ippressed by lesions in units to which they do connect. Units ith very widespread connectivity have a much less variable esponse to lesions. They are less affected by lesions in the ructures to which they connect, because may of their inputs emain intact, but they are also more sensitive to the indirect etwork-mediated effects of lesions in structures to which they o not connect directly.

he resulting decrements in activity in the structures epresented. This simulation revealed three effects. First, he effects of lesions propagated to other structures to hich the lesioned structure was not directly connected, as well as to those structures in receipt of direct projections from the lesion site. Second, the impact of a lesion on activity in all the other structures in the network depended on the number and strength of connections in which the lesioned structure participated. Third, the vulnerability of structures to lesions elsewhere in the network again depended on the number of connections of a structure. Structures with profuse connectivity were affected by lesions in many other structures, but the magnitude of their suppression did not depend greatly on the precise location in which the distant lesion was made. Structures with relatively few connections were greatly affected by lesions in the few structures from which they received connections, but were much less sensitive to lesions in the many structures with which they did not connect.

These three effects are each unsurprising. Connection diagrams themselves promote a recognition of the plethora of pathways through which information could be conducted. Similarly, the dependence of the impact of a lesion, and the dependence of vulnerability to distant lesions, on the richness of the connectivity of structures could be apprehended from first principles. However, some of these effects do not appear to have been considered in the context of the inferences that can reliably be made about the functions of brain structures from the effects of their lesions on behaviour. The question arises: Could conventional inferences about the effects of brain lesions reliably determine the functional roles of structures in a network that behaves like that simulated in the previous section?

The question of what constitutes reliable evidence for an imputation of a function to a structure has been treated by neurologists and behavioural neuroscientists (e.g. Dean 1982; Damasio & Damasio 1989; Grobstein 1990; Luria 1973; Teuber 1955). In general, these treatments have tended over time towards increasingly great caution in what can validly be inferred from the effect, or lack of effect, of a lesion on behaviour. Typically they have focused on the inferential adequacy or otherwise of varieties of dissociation of function revealed by lesions, and we examine these dissociations in the context of the behaviour of the thalamocortical model below. However, inferences about which part of the brain does what made from data about behavioural lesion effects are to be distinguished from the inferences made in a different enterprise from somewhat similar data. Aspects of what can be deduced from the effects of lesions about information processing and other functional models have also been discussed extensively by neuropsychologists (e.g. Jones 1983; Shallice 1988). Since the aim of this neuropsychological work is mainly to dissociate functional models and not to impute functions to particular structures in the brain (Shallice 1988), it presents a different problem to that of imputing function to structure on the basis of the effects of brain lesions, and we do not treat it further here.

(a) *Indirect effects and diaschisis* ('*action at a distance*')

Indirect effects, mediated by multiple routes through multiple structures, are a feature of a relationship between cortical connectivity and the patterns of spread f disinhibited cortical activity (Kötter & Sommer, this ssue). Indeed, removing these indirect interactions from he computations reduces the goodness of statistical fit etween connectivity and activity spread (Kötter & ommer, this issue). Here, very similar integrative nechanisms suggested that indirect interactions should lso arise from, and relay, activity decrements resulting rom lesions. Activity decrements in a structure, arising rom reduced inputs from distant lesioned or inactivated tes, could affect the mediation of the structure's nformation processing functions (e.g. Hilgetag, Burns,)'Neill, Scannell & Young, this issue). Should this local nformation processing function be vital to the performance of a behaviour, lesions at distant sites could therefore affect the behaviour by these indirect means, even vhen they play no direct role in mediating it. Hence, ction at a distance', or diaschisis, a concept once much sed among neurologists (e.g. Monakow 1910, 1914; uria 1973), but which appears to have fallen almost out f use, should be a fairly general property of brain etworks.

In the context of inferences from lesion effects, it is a trong temptation to believe that an experimentally nduced lesion causing a decrement in a behaviour does o directly through the impairment of the information rocessing functions of the lesioned structure. However, ndirect network-mediated effects, if present, suggest that is unsafe to assume that a lesion has its detrimental ffect on behaviour by virtue of the effect of the lesion on rocessing local to the lesion site—or even on processing n the structures to which the lesion site is directly onnected. Evidence that sites distant to the lesion are nimpaired would be required in addition to the lesion ocation and the functional deficit for the impaired funcon to be imputed to the lesioned structure. Plainly, this dditional information could not be derived without aining further information on processing elsewhere. onclusive proof for the imputation would require an xhaustive search through all other possible brain structures, since the inference takes the form of argument by xclusion. Hence, if there is any propagation of activity ecrements from a lesion through the network sufficient o degrade information processing elsewhere, the implicaon is that the loss of a behavioural function following a esion cannot be adequate to infer that the lesioned strucure was involved in mediating the degraded function. lore directly empirical considerations led to the same onclusion (Grobstein 1990).

In a similar vein, the propagation of lesion effects way from the lesion site implies that the lesioned etwork will be inequivalent to the intact network, even eaving aside the differences of processing in the lesioned tructure, and considerations of possible plastic change lsewhere. This also restricts what can be inferred from he survival of a particular function following a lesion. etention of the function plainly suggests that the emaining structures and circuitry are sufficient to nediate the behaviour in the lesioned animal. But the nequivalence of the lesioned and intact networks uggests that it would not be justified to infer that the on-lesioned structures are sufficient in the intact system see also Grobstein 1990). Similarly, this inequivalence urther suggests that it would not be justified to infer that the lesioned structure did not mediate in the intact system a function that remains after the lesion. For example, it could not be validly inferred from the preservation of aspects of colour vision after a lesion of V4 that V4 did not mediate these same aspects of colour vision in normal vision in an intact animal prior to the lesion (cf. Heywood *et al.* 1995).

These foregoing considerations of the validity of inference arise from the propagation of the effects of a lesion to distant elements of the brain's network. In the next section, we turn to the specific issue of single dissociations of function.

(b) *Single dissociation*

Lashley (1952) and Teuber (1955) raised the question of whether an apparently specific deficit arising from a lesion can be sufficient proof that the deficit is actually specific. An apparently specific deficit could indeed arise from the loss or impairment of a specific process and processor, but the possibility that the deficit arises from some more general impairment could not be ruled out by a single dissociation of this kind (Teuber 1955). Hence, initial questions about the adequacy of single dissociations of function arose from suspicions that such results could not rule out more general deficits that could explain experimental results just as well. However, the nature of the deficit, and the experimental circumstance in which it appears, determine to an extent the plausibility of alternative, non-specific, explanations for it. Some results are easier than others to challenge in this way. For example, it would be easy to invoke any of a variety of general impairments to explain the loss of food-acquisition behaviour following a lesion. It may be harder to explain in non-specific terms the loss of orientating behaviour towards food items presented in the visual field contralateral to a cortical lesion when this is accompanied by intact orientating to the ipsilateral hemifield and by control conditions that rule out lack of comparison behaviour, a failure of comprehension of the testing situation and differences of the training set (e.g. Lomber & Payne 1996). Hence, competing non-specific accounts for particular deficits might be ruled out or ameliorated by careful experimental design, as for other methodologies, all else being equal.

A second defect of single dissociations as a basis for imputing function to structure, however, was a concern that some functions may be mediated by processors that are more sensitive to damage anywhere in the system (e.g. Teuber 1955). A behavioural deficit apparent after a lesion could be an example of the decrement of a vulnerable processor by a lesion in a structure that itself has no information-processing role in mediating the behaviour, or it could be evidence for an interdependent hierarchy of function in which the lesioned site plays a role, rather than evidence for a localization of the function (Teuber 1955). These possibilities cannot be ruled out by a single dissociation of function, even with very careful experimental design, since they advert to aspects of the internal organization of neural systems that are impossible to control externally. These concerns have led to great caution in making inferences about the localization of function from instances of single dissociation (e.g. Grobstein 1990; Teuber 1955). It is now widely recognized hat a loss or deficit in a behavioural function that follows lesion in a particular structure does not imply that the tructure was involved in mediating the function Grobstein 1990).

Both of the effects that arise in our simulation from the ifferent numbers of connections possessed by different tations suggest that this reticence about single dissociation well advised. Some structures were straightforwardly nore vulnerable to lesions elsewhere than others. Should n experimenter have the misfortune to take an interest in behavioural function mediated by one or more especially ulnerable processors, and the further misfortune not to sion one of these implicated structures, a deficit in the ehaviour in a single dissociation would immediately lead the imputation of the function to the wrong station. An xperimenter with uncommonly greater luck might make he right imputation, but the right and wrong cases cannot e discriminated without further information. Single issociation is therefore capable of correct imputation: the roblem with its reliability as a basis for inference is not a asic logical incapacity, but that one cannot know without ther information that the inference is correct. Hence, ifferential vulnerability of brain structures strongly uggests that a single dissociation of function is not reliable vidence for the imputation of a function to a structure, as oted by Teuber (1955).

Earlier discussions of single dissociation, however, do ot appear to acknowledge the other factor that arises rom differential connectivity: the differential impact of sions on the network. The simulation showed, unsurrisingly, that lesions of structures emitting relatively arge numbers of connections affected structures lsewhere in the network more than did lesions of egions with few connections, and that direct connecions were particularly effective in propagating decrenents to stations with few connections. This provides nother way in which luck could enter the imputation of unction from a single dissociation. Experimental lesions structures other than that mediating the function eing tested could be made in regions with a paucity of onnections and no direct connection to the processors nediating the function, so avoiding misattribution of the unction to them. But such a lesion made in a richly onnected structure, or in one emitting a direct connecion, might reduce activity in the mediating processors ufficiently that the behavioural function would be nputed incorrectly to the wrong richly connected or irectly connected processor.

These considerations suggest that single dissociation is ot a reliable means of imputing function to structure in he brain, because it can easily give rise to incorrect ttributions. The differential vulnerability and impact videnced by the simulation echo concerns that have long een credited in neurology and behavioural neuroscience. hese disciplines have consequently developed a more laborate basis for inference about the roles of brain tructures. Double dissociation now represents for many he 'gold standard' for inference and has been considered provide 'conclusive proof' (Teuber 1955). The next ection considers the validity of double dissociation as a neans of imputing function to structure in the context of he effects of lesions made evident by the simulation of he thalamo-corticocortical network.

(c) *Double dissociation*

Following Teuber (1955), an example of double dissociation is that tactile discrimination can be disturbed by some lesion without loss on visual tasks, to a degree of severity comparable to visual deficits arising from a different lesion, which lesion causes no loss on the tactile task. Hence, more generally, double dissociation is the case when function 1 is disturbed by lesion A and not lesion B, while function 2 is disturbed by lesion B and not lesion A. Inference from double dissociation offers much stronger evidence that the two functional deficits are specific than does single dissociation (Teuber 1955), but it has not been prescribed principally to impute functions to structures, despite having very frequently been used to do so, particularly in recent years (e.g. Ennaceur *et al.* 1997; Hunt & Aggleton 1998; Killcross *et al.* 1997; Ragozzino *et al.* 1998; Sahakian *et al.* 1995; Selden *et al.* 1991).

Experimental studies already suggest that imputations of function from double dissociation require caution. Consider, for example, the fact that orientating to the left visual hemifield is abolished by right parietal cortex inactivation, but is unaffected by left parietal inactivation; and that orientating to the right visual hemifield is abolished by left parietal inactivation, but is unaffected by right parietal inactivation (Lomber & Payne 1996). This pattern of results represents an unequivocal double dissociation of function between left and right orientating behaviour, made all the clearer since these effects are reproduced in the same animal by reversible inactivations (Lomber & Payne 1996). The abolition of orientating contralateral to the inactivated parietal region, and the fact of the double dissociation across the midline, could be taken to suggest straightforwardly that right parietal cortex mediates orientating to the left, while left parietal cortex mediates orientating to the right. However, bilateral inactivation of both sites simultaneously results in orientating to both visual hemifields being paradoxically restored (Lomber & Payne 1996), indicating that other systems in the bilaterally inactivated circumstance are capable of mediating apparently normal orientating behaviour (Hilgetag *et al.* 1999). Hence, double dissociation here suffers the same inferential uncertainties that attended imputations of function from single dissociation, and indeed inherits these same uncertainties from the single dissociations that combine to form the double dissociation. In this case, neither of the abolitions of contralateral function allow reliable inference that the inactivated structure mediated the abolished function; it cannot even be stated with certainty from these results that the two parietal sites were involved in orientating function, since distant effects of their inactivation on processors that were involved may have been responsible for the deficits (e.g. Hilgetag *et al.* 1999). We note that these uncertainties about double dissociation would not have been emphasized without the startling and paradoxical effects of multiple inactivations, made apparent by careful studies of this system (e.g. Sprague 1966; Lomber & Payne 1996; Wallace *et al.* 1989, 1990).

What could be concluded about the validity of double dissociation from the effects observed in the simulation of the thalamocortical model? Interactions between differential vulnerability and impact are of particular interest, since it is possible that these two factors might conspire to able 1. *A table of qualities related to the expected categories of severity for a variety of possible lesion and processor combinations*

he qualities are generated by interactions between the differential impact of lesions and differential vulnerability to lesions, oth of which effects were related to the different numbers of connections possessed by structures in the thalamo-corticocortical mulation (see §2). We consider the simple case of effects on two notional processors, one a richly connected (RC) station and e other a less-connected (LC) one. The two processors mediate different behavioural functions. We assume that lesion of either ation would abolish the function being performed there (effects of severity XXXX). For lesions made elsewhere in the network an these two processors (i.e. 'misses'), the combinations of impact of such a lesion and the vulnerability of the processor to such lesion are expressed in the other qualities. For example, the LC processor is relatively invulnerable to lesions made in stations nconnected to it (i.e. MISS–INDIRECT cases), and lesions in some other LC station have a relatively modest effect on ructures elsewhere in the network. Hence, LC–MISS–INDIRECT produces an effect of low severity, X. The RC processor is latively more vulnerable to lesions made elsewhere, and so this combination of lesion and processor produces an effect of verity XX. Similarly, lesions made in RC structures have a greater impact on other structures and so produce effects of greater verity than those in LC stations. The quality of severity of every combination of processor and lesion can be derived in the me way from combinations of vulnerability and impact. The categories RC–HIT on a LC processor and LC–HIT on a RC rocessor do not exist. The consequences of these contingencies for inference using single and double dissociation are described in e text.)

	lesions of RC stations			lesions of LC stations		
	RC–HIT	RC–MISS–DIRECT	RC–MISS–INDIRECT	LC–HIT	LC–MISS–DIRECT	LC–MISS–INDIRECT
C	XXXX	XXX	XXX	—	XX	XX
C	—	XXX	XX	XXXX	XXX	X

roduce effects of unsuspected severity in surprising ations. To explore this issue initially, we considered two ifferent behavioural functions, one delegated to a richly onnected, and the other to a less-connected station. Jsing the results of the simulation as regards the vulnerbility to, and impact of, lesions to structures with these onnectional properties, we constructed a contingency able to show the quality of the severity of effects that ould be expected for each combination of lesion and unction.

The rows of table 1 give the quality of the effects on the wo different behavioural functions of different lesions. he top row corresponds to a function mediated by a ichly connected (RC) processor and the lower row to ne mediated by a less-connected (LC) processor. esions can be made in RC or LC stations; and the esions can be either direct hits on the processors oncerned or made elsewhere (as in accidental misses or ontrol lesions). For lesions made elsewhere in the etwork, there was a marked difference in the simulation the effects of lesions made in structures directly onnected to LC structures, when compared to the effect f lesions to structures not directly connected to them. his difference is represented by the MISS–DIRECT .e. a miss lesion made in a structure directly connected the processor mediating the function) and MISS–NDIRECT (i.e. a miss lesion in a structure not directly onnected to the processor mediating the function) ategories. Complete abolition of the function could be gnalled by XXXX qualities, severe degradation by XX, moderate or noticeable deficit by XX and minor r insignificant effects by X qualities.

We consider a threshold for determining a significant ehavioural decrement in the function that lies between ffects of strength XX and XXX. This threshold can be aised or lowered, for example by using a more or less ensitive behavioural test, or by altered statistical criteria. Iowever, a lower threshold (i.e. between effects of everity X and XX) would render functions mediated by a RC processor impossible to localize, because its function would always be disrupted significantly by any lesion anywhere. Hence, there could be no double dissociation in this case, because the RC function would always be disrupted. Functions carried by an LC processor could also not be localized in this case because only lesions in less-connected structures not connected to the LC processor would yield informative preservation of the LC processor's function and a process of elimination could not therefore be conducted. Empirical results show that double dissociations do occur and so a lower threshold for deciding whether a significant behavioural deficit has occurred is unrealistic. Conversely, complete abolition of a behaviour is seldom a requirement for a dissociation to be claimed experimentally, and so a higher threshold (i.e. between effects of severity XXX and XXXX) is also unrealistic.

Does the table of severities in table 1 provide a basis for the correct assignment of functions to structures using double dissociation? Consider two lesions, one made in the RC processor that mediates function 1, the other made in the LC processor that mediates function 2. The first lesion (RC–HIT) abolishes function 1 (effect of severity XXXX). Concomitantly, if the RC processor is assumed to be unconnected to the LC processor, the same lesion would also constitute a miss in a richly connected structure unconnected to the LC processor (RC–MISS–INDIRECT), yielding a non-significant effect of severity XX. The second lesion (LC–HIT) abolishes function 2 (XXXX), but does not significantly degrade function 1 (LC–MISS–DIRECT or LC–MISS–INDIRECT: both effects of severity XX). Hence, lesion A degrades function 1 but not function 2, while lesion B degrades function 2 but not function 1, constituting a double dissociation. In this circumstance, function 1 would be correctly imputed to the RC processor that mediates it and lesion of which abolishes function 1. Similarly, function 2 would be correctly imputed to the LC processor that mediates it and lesion of which abolishes function 2.

Exactly analogous contingencies can be explored for wo different functions mediated by two different LC rocessors or two different RC processors. In the case here the two LC processors mediating the two functions re unconnected, a double dissociation can again be erived that correctly ascribes the functions to the two rocessors, provided that the lesions are made in the orrect processors. However, in the case that the two LC rocessors are directly connected, both lesions would ignificantly degrade both functions, because of the relaively high impact on an LC processor of a lesion made in nother structure directly connected to it (LC−MISS−DIRECT), and hence no double dissociation might be erived as a basis on which to impute function to a rocessor. A similar problem could attend imputations of ifferent functions to two different RC processors. The igh impact on the network of a lesion in an RC station, nd the vulnerability of a function-mediating RC rocessor to lesions elsewhere, mean that significant egradation of both functions could follow from any esion of an RC structure. Hence, double dissociations ight be expected to be more difficult to demonstrate for hese combinations of processors and lesions, and so there ould be greater difficulty in using double dissociation to npute the functions to structures in these cases. Also, the reater impact of lesions made in structures directly onnected to processors mediating a function should ender it more difficult to generate clear double dissociaions. This might make it more difficult to impute ifferent functions to directly connected processors by hat form of inference, assuming no gross difference in he connectivity of the two processors to the rest of the etwork.

Does the table of severities in table 1 provide a basis or the mistaken assignment of functions to structures in ases of unequivocal double dissociation? Consider again he circumstance that function 1 is mediated by a RC rocessor, and function 2 is mediated by a LC processor. onsider further a lesion made in a RC structure that oes not mediate behavioural function 1 (as in the RC−ISS−DIRECT and RC−MISS−INDIRECT columns). ecause of the large effect on the network of lesioning he RC structure, and the vulnerability of the RC rocessor itself to lesions anywhere in the network, the esion could severely degrade the function (effect of everity XXX). The same lesion, if the RC structure and he LC processor are unconnected, does not decrement he LC processor's function significantly (LC−MISS−NDIRECT: XX). A different lesion, making a direct it on the LC processor mediating function 2, will egrade the LC function (XXXX), but it is also a LC−ISS−INDIRECT (XX) for the RC processor, and it oes not decrement the RC function significantly. Hence, esion A degrades function 1 while leaving function 2, nd lesion B degrades function 2 without significantly ffecting function 1. These lesions therefore generate an nequivocal double dissociation of function and an nequivocally incorrect imputation of function to ructure: function 1, mediated by the RC processor, is istakenly imputed to the wrong RC structure. Similar xamples of defective inference can be derived from ases in which the ascription of function to the RC rocessor is correct, but the imputation of the LC processor's function is incorrect; in which functions are mediated by two LC processors, a lesion is made in a station connected to one but not the other processor, and one or both functions misascribed; and so on.

These considerations suggest that counter-examples, in which incorrect imputation of function to structure is made, can be demonstrated readily for both single and double dissociation using simple principles of likely interaction between brain structures. Double dissociation appears therefore to suffer similar problems of unreliability as have long been recognized to diminish the significance of single dissociations: while inferences from double dissociations can correctly ascribe functions, they can also yield incorrect imputations, and only further information can discriminate correct from incorrect cases. Hence, if the simple propagation effects of lesions derived from the simulation in 2 obtain in the real brain network, neither single nor double dissociation derive reliable information about the functions mediated by brain structures.

4. CLARIFYING 'FUNCTION' AND A FRAMEWORK FOR INFERENCE

The considerations in § 3 suggest that conventional inference from single and double dissociation may be defective as a means of determining reliably what different parts of the brain do. On the other hand, most of the many imputations of function to particular brain structures derived from the effects of lesions have been borne out to some extent by subsequent research with a wider variety of methodologies. Testing the behavioural consequences of brain lesions suffers from well-known technical problems in inactivating structures and in testing the behavioural outcomes in a sufficiently fine-grained or insightful way (e.g. Grobstein 1990). But these technical difficulties are in many cases tractable, and reliable information derived from these methods should be very valuable in understanding how the brain mediates behaviour. We were motivated, therefore, to try to develop reliable inference for imputing function from this kind of data. However, as pointed out by Teuber (1955) 'no degree of refinement of ... technique can substitute for clarity of concepts referring to structure and function. ... Unless we work on our concepts, the accumulation of facts will hinder rather than help'. Accordingly, this section re-examines concepts invoked by the search for structure–function relationships, in the pursuit of greater clarity, before going on to suggest a more formal framework for inference.

Making a lesion in a brain structure and then testing for a behavioural change is a prototypical example of a methodology for seeking structure–function relationships. As we have described, structure–function relationships presently remain rather opaque at many scales of the nervous system. However, we do not believe that this opacity arises primarily from deficiencies in current understanding of structure as it derives from neuroanatomical data. There are many uncertainties in neuroanatomical parcellation and connectivity (e.g. Colby & Duhamel 1991; Stephan, Hilgetag, Burns, O'Neill, Young & Kötter, this issue; Young *et al.* 1995; Hilgetag, Burns, O'Neill, Scannell & Young, this issue) but these are, in he main, experimentally tractable problems, rather than rising from failures of clarity in the concepts being pplied. Conversely, there seems to us a lack of clarity in hat is meant by 'function'. This confusion makes connecons between brain structure and 'function' difficult to pecify rigorously. We discriminate at least five different, artially overlapping senses of function in frequent but onflated use, and think it instructive to try to disentangle hese different senses of function.

Function appears to be applied in at least the following ifferent senses: the evolutionary biological sense, of funcon as survival function, f_e; function as a discrete local roperty, f_l; function in the context of the network, f_c; unction in the sense of the function of the global nervous ystem, as in its behaviour, f_g; and function in the formal ense of a mathematical mapping between input and utput, f_m. These different senses of function are now iscussed in turn.

(i) Function (evolutionary, f_e). This sense of function is concerned with the presumed evolutionary fitness benefits conferred by particular structures. We might ask of a structure, for example, what advantage it gives its bearer. In this sense, the function of some structure or organization is related to the selective advantage bestowed, eventually in terms of enhanced survival and reproduction, relative to an individual with a different structure or organization. A function (f_e) of the tectospinal tract might thus be to support differential survival and reproduction through improved eye–claw coordination, given that its relative size correlates with predatory habits (Barton & Dean 1993). Similarly, a function (f_e) of the parvocellular compartment of the lateral geniculate nucleus might be to support differential survival and reproduction through improved ability to select ripe fruit using colour vision (Barton 1998). This sense of function, in terms of fitness benefits or survival function, should converge with some of the senses of function below. This is because neural systems are biological mechanisms, and the only known way for biological mechanisms to come about is by selection acting on variability. Hence, characterizing function in relation to the selection pressures that have acted and act to adapt neural systems should relate closely to more causal aspects of function (e.g. Cosmides & Tooby 1995) since, in general, neural systems are what selection has caused them to be and they do what selection pressures require of them.

(ii) Function as a discrete local property (f_l). This sense of function concerns the function of a component of a system when considered as an isolated element, disconnected from the system in which it is normally embedded. Consider, for example, the printed circuit board of a radio. If we were to clip out a capacitor from it, we might say that the capacitor's function is to store charge. This description of its function might be wholly different when made in the context of the rest of the circuitry (see below). In the same way, if we were to consider a single neuron in isolation from the networks that embed it in the brain, we might say that its function (f_l) is to integrate its inputs and produce an output spike stream contingent on those inputs.

(iii) Function (in context, f_c). This sense concerns the function of a component in the context of the network that surrounds it. Hence, if we were to resolder the capacitor whose f_l was to 'store charge' back into the radio, we might now say of it that its function is to act as a high-pass filter to aid tuning into different radio stations. In the same way, the function of a brain component in this sense can be understood only in the context of the wider structure of the network of which it is a part. Hence the function (f_c) of V4, for example, is determined by the nature of its inputs, its internal computations, its extrinsic connectivity and the nature of the rest of the network, which determines the role of this structure in the global information processing economy.

(iv) Function (global, f_g). Function in this sense relates to the behaviour of the whole animal. We might say that orientating to food items in the left hemifield is a function (f_g), and that orienting right is another function (f_g). These functions could be fairly complex, since this sense of function concerns anything an animal can do. Many such functions can readily be characterized in terms of inputs, internal computations, including the retention of information over time, and behavioural output.

(v) Function (formal–mathematical, f_m). This sense of function is the literal one, concerning the mapping of inputs on to outputs and the transfer function involved in this process. Thus one might treat the function of V1 by examining the mapping of its inputs from the LGN, V2, V3, V3A, V4, V4t and MT (V5) on to its outputs to the LGN, V2, V3, V3A, V4, V4t and MT. Similarly, one might treat the global function of the whole animal, for example, during psychophysical performance or an experiment on orientating behaviour, by examining the mapping between input and output. Indeed, this sense of function could apply to the whole animal, and any processor, set of processors or subprocessor within the brain, provided that the input, mapping and output of each function are sufficiently well characterized as to be capable of mathematical specification.

Which of these different senses of function, or which explicit combinations of them, are the most useful in considering brain structure–function relationships? We turn first to the usefulness of function in the discrete, local sense (f_l). To explore the relationship between structure and f_l for a component, the component must be capable of being considered both structurally and functionally discrete: that is, there must be an interface external to the component at which it can be separated from the remainder of the system. Consider, for example, an electronic circuit board, in which the components (e.g. chips) are perfectly discrete and their extrinsic connectivity is just that—extrinsic. Solder can be applied at the interfaces between components, and between the components and the circuitry, to join them to the rest of the system. In the case of neuronal microcircuits and all more molar structures in the nervous system, however, there is no external terface at which one could imagine a neural solder being pplied. Extrinsic projections, such as corticocortical rojections from neurons with distant cell bodies, reach ight into the circuits themselves, and in that way form an trinsic part of them. Neuronal circuits, including micro-ircuits, are themselves formed in part by synapses made y cells with distant cell bodies, or are otherwise intiately affected by distantly derived factors. A consequence f this feature of brain organization is notable in modelling udies: modelling small patches of isolated cortex, for xample, always involves setting arbitrary boundary onditions that do violence to the actual processing archiecture in the real brain, simply because a substantial umber of the synapses in any volume of tissue are made y neurons with distant cell bodies. Hence, in the case of he brain, there may be little benefit in defining a local nction for a multineuronal component, since removal of s ‘extrinsic’ connectivity renders a different component. upra-neuronal structures are not discrete, and so do not ave functions in this sense. Above the level of the single euron, therefore, f_l may not be a useful concept.

Function in the evolutionary sense (f_e) seems more eadily applicable to brain structures and systems. During he past two decades strong progress has been made in nalysis of the evolutionary ecology of animal behaviour .g. Krebs & Davies 1978, 1997), and results in that area rovide a functional (f_e) framework that will aid underanding of the causal aspects of function, which are of rimary interest to neuroscience (Cosmides & Tooby 995). However, most attempts to relate evolutionary nction to neural structure have been focused on the ensory periphery and on relatively simple aspects of nimals’ ecology. While the adaptation of retinal photoigments to the spectral properties of important fruit od items has been characterized (e.g. Osorio & orobyev 1996), for example, much less is known about ow central neural systems are adapted to mediate daptive behaviour in foraging, mate choice, antiredator vigilance, sexual signalling or the many other spects of animals’ ecology which are now known to be nder strong selection pressures. Part of the problem in elating evolutionary function to central brain structures that evolutionary studies have largely been undertaken rather many species for which there is relatively little etailed neuroanatomy or neurophysiology and, converly, that detailed neuroscientific investigations have argely been undertaken in a small number of species that ave not always formed the primary foci for studies in ehavioural ecology (but see Turner & Bateson 1986). ence, detailed functional information on a species’ cology and behaviour is often accompanied by relatively rude neuroscience, and vice versa. Bringing the potenially great information from evolutionary ecology to ear on brain systems will require a concentration of both ypes of study on the same species. Presumably, these ould initially be the laboratory species on which there is lready a wealth of neuroscientific information, since this formation takes much longer to acquire than does inforation about ecology and behaviour. At present, owever, the salience of evolutionary considerations to nderstanding brain structure–function relationships is mited by this lack of concordance in the species being udied.

Neuroscientific discussions of the functions of multineuronal groupings, such as cortical areas or thalamic nuclei, implicitly use a sense of function closest to f_c (function in the context of the system), although the term is most often used with negligible recognition of the dependence of this concept on distributed and contextual factors, such as extrinsic connectivity, the organization of the networks in which structures are embedded and the dynamic system-wide context in which a structure’s computations are performed. The concept of the function of a brain component being understood in the context of the nature of its inputs, its internal computations, its extrinsic connectivity, and the structure and dynamics of the rest of the network, which eventually determine the role of the structure in behaviour is, for the present purposes, sufficiently precise to allow a formal specification. Indeed, the literal sense of function, as a mathematical mapping between inputs and outputs, can be applied to the function of each constituent structure by employing terms for inputs, the mathematical mapping between these inputs and a structure’s outputs, and for interactions between all structures as specified by their connectivity. Similarly, global function, f_g, pertaining to the behaviour of the whole animal, can also often be characterized in terms of inputs, behavioural output and the mapping between them.

These more explicit specifications of what is meant by function permit a more formal framework for exploring the relationships between brain structure and function. Our aim remains to derive valid rules of inference for imputing function to structure in the brain. Ideally, such rules of inference should be derived from a mathematical treatment of the functions of brain structures, the function of the whole animal as manifest in its behaviour and the relationship between these functions. For the present, we formulate the problem as follows. Consider a whole-animal function, f_g, such as orientate left to food items presented in that hemifield. This function could be captured formally as a mapping between stimuli presented to the left visual field and motor output which moves the animal to the appropriate location. We then consider any such global function f_g to be delegated among the processors in the brain in such a way that some set of processors’ functions (f_c) are sufficient to generate the global mapping observed. Each processor’s function f_c could also be captured formally as a mapping between its inputs and outputs in the context of the connectivity and dynamics of the system. Each structure’s function will bear a relationship to the global system function, which could be captured quantitatively by the loading of each structure’s function on the global system function being tested. The problem of imputing function to neuroanatomical structures on the basis of the effects of brain lesions then becomes the task of discovering the loadings of structures on the global function through observations of the effects on f_g of lesioning the network; that is, the problem of determining the loadings of structures’ f_cs on f_g from lesion-generated changes in f_g.

Loadings in this framework estimate the quantitative importance of a structure that bears a particular loading for mediating that function. A high loading signifies that a particular structure is important to mediating the function.

n the limit case, a single processor might mediate a global ınction by itself, so possessing a loading of 1.0. In this ase, its inputs, computations and outputs would be ıfficient for the mapping between external input and ehavioural output to be undertaken locally by the rocessor (i.e. $f_c = f_g$). This would require the processor to ossess the correct connectivity, sufficient activity and ppropriate information, for the following reasons. A rocessor that possessed the appropriate information and ould broadcast it with sufficient activity could not mediate ıe function if its outputs were not directed to the correct ownstream structures (e.g. motor structures). Hence onnectivity is a determinant of the importance of a ructure in mediating a behavioural function. Similarly, a rocessor with inputs sufficient to acquire the necessary ıformation, and outputs bearing correct information and irected to appropriate structures, could not mediate the ınction if its activity were so low that its output signal ould not affect processing in its targets. Consequently, ctivity is a determinant of the importance of a structure ı mediating a behavioural function. Furthermore, a rocessor with appropriate connectivity from inputs and to utputs, capable of broadcasting its activity with sufficient ain to affect processing in its targets, could not mediate ıe function if its outputs were void of information or were isinformative. Information is thus a determinant of the nportance of a structure in mediating a behavioural funcon. Hence, at least three factors determine the loading of structure's function on the global function. These loadıgs are scalar quantities, however, and capture only the nportance of a structure to the mediation of a particular ehavioural function. Loadings do not capture how or hat the processor contributes to the mediation of a behaiour. We assume that behavioural functions are not often ıediated equipotentially by very many different structures ith roughly equal low loadings, and we note that the ame structure can readily possess different loadings on ifferent global behavioural functions.

5. A WORKED EXAMPLE OF INFERENCES FROM BRAIN LESIONS

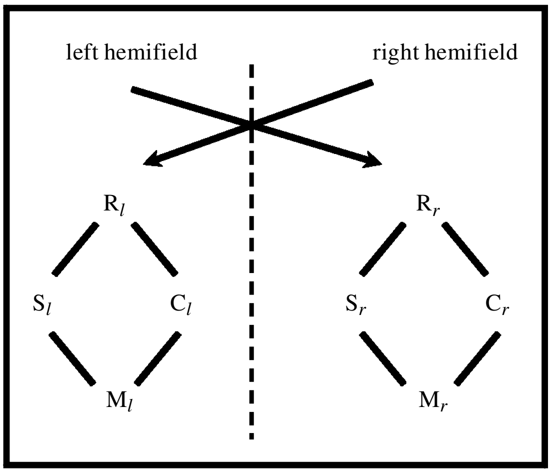
In §4, we attempted to clarify useful concepts of funcon. Using this clarification, we set out a more formal pproach to imputing function to structure on the basis of rain lesions. The present development of this framework shown without mathematical formulation, and we now ırn to a worked example of the use of this framework to ıow more practically how it could be applied. To do this e examine inferences about the locations of function elating to a simple network that reproduces intact, esioned and paradoxically restored orientating behaviour after Hilgetag *et al.* 1999), and seek to determine whether ıe task of discovering the loadings on the behavioural ınction of particular structures by lesioning the network tractable in this simple system.

Neurologically intact cats can direct their attention to ood items presented anywhere in their visual fields. Cats ith unilaterally lesioned or inactivated parietal cortex ail to orientate to visual stimuli appearing in the ontra-lesional hemifield (Sprague 1966; Payne *et al.* 996*a*,*b*). The same failure is apparent after unilateral esion, or inactivation, of the superior colliculus (Lomber & Payne 1996). However, Sprague found that the visual hemi-extinction induced by damage to one posterior cortex in the cat can be paradoxically reversed by subsequently damaging further structures, in addition to the primary lesion. Orientating can, for example, be restored by secondary lesions in the superior colliculus on the contra-lesional side (Sprague 1966). Similarly, paradoxical restorations of function after bilateral inactivation of the cortical sites and bilateral inactivation of the colliculi occur, demonstrating that subsequent inactivation at the same level as the primary lesion can restore performance (Lomber & Payne 1996). These results form a complex, and somewhat perplexing and counter-intuitive, set of effects, which are nevertheless experimentally robust.

We have previously developed a simple model based on known connectivity to account for these perplexing results (Hilgetag *et al.* 1999). The details of the model help to explain, in addition to the results above, the slower and more partial restoration of function that follows section of the commissure of the superior colliculus and the failure to restore orientating function to the far periphery following lesions that otherwise restore function (Hilgetag *et al.* 1999). An even simpler account, however, is sufficient for intact orientating, unilaterally lesioned impairments in orientating, the paradoxical restoration of function in the Sprague paradox, and the paradoxical restorations in both the cortical and collicular Payne–Lomber paradoxes (see below).

Consider a system in which two bilateral systems exist, one cortical and the other subcortical, and in which balanced competition between sides is the basic principle of operation. In the intact system, a stimulus presented to one visual hemifield produces greater activity in both cortical and subcortical structures contralateral to it. This greater activity on one side engages motor output and unilateral orientating behaviour is emitted appropriately. Any single unilateral lesion so diminishes activity on that side that, even with the benefit of stimulus-related activity, activity on that side is insufficient to overcome baseline activity on the other. Hence, no appropriate capture of motor systems takes place, and appropriate orientating is abolished. Any pair of contralateral lesions, however, will render a bilaterally balanced system. Any such system can be unbalanced by stimulus input, and so capture motor systems appropriately, reinstating correct orientating. For example, bilateral inactivation of the colliculi yields a balanced bilateral system comprising the two parietal structures and restored orientating as in the Payne–Lomber collicular paradox (Lomber & Payne 1996; Hilgetag *et al.* 1999). Bilateral inactivation of the two parietal cortices yields a balanced bilateral system comprising the two colliculi and restored orientating as in the Payne–Lomber cortical paradox (Lomber & Payne 1996; Hilgetag *et al.* 1999). Similarly, unilateral inactivation of parietal cortex, together with inactivation of the colliculus contralateral to it, yields a balanced bilateral system comprising one cortical and one subcortical processor and restored orientating as in the classical Sprague paradox (Sprague 1966).

Figure 3 provides a diagrammatic representation of the simple network required to implement these effects. Consider that the network implements two global



igure 3. A notional model sufficient to account for some spects of orientating behaviour and the changes in this behaiour after lesions. The model is abstracted from a detailed athematical model (Hilgetag *et al.* 1999), which is itself ased on anatomical structures and connections thought to be volved in visual orientating in the cat. It is presented only to otivate the discussion of how the loadings of the processors n the global behavioural function of the system might be ecovered through observing the effects of lesions (see text). he model contains two bilateral systems, composed only of e 'colliculi', S_l and S_r, and the 'cortical' structures, C_l and $_r$. In the baseline state, there is a balance of activity between e two sides. Activity related to stimuli in a visual hemifields relayed to both the contralateral structures. The side with reater activity captures the motor plant and behaviour is mitted toward the hemifield contralateral to the more active de. The effects of lesions on this simple network are set out in e main text.

unctions, orientate left $f_{g\_l}(x)$, and orientate right, $_{\_r}(x)$, on a sensory input, x. Stimuli can either be resented on the left, $x=l$, or on the right, $x=r$, or can be bsent or central. Each global function has two discrete utput states. The outputs of $f_{g\_r}(x)$ are orientated right, or do nothing, null. The outputs of $f_{g\_l}(x)$ are orientated eft, l, or do nothing, null.

$$\left.\begin{aligned} {}_{g-l}(x) &= \begin{cases} l \text{ when } x = l \\ null \text{ when } x \neq l \end{cases} \\ {}_{-r}(x) &= \begin{cases} r \text{ when } x = r \\ null \text{ when } x \neq r \end{cases} \end{aligned}\right\} . \qquad (3)$$

onsider that the left and right colliculi and cortices have omponent functions that contribute to the global funcions, $f_g(x)$, equally in the intact state. Here the component unctions, f_c, of the right cortex and colliculus, and left ortex and colliculus are given by $f_{c\_cr}(x)$, $f_{c\_sr}(x)$, $f_{c\_cl}(x)$, nd $f_{c\_sl}(x)$,

$$\left.\begin{aligned} {}_{g\_l}(x) &= f_{c\_sr}(x) + f_{c\_cr}(x) + f_{c\_sl}(x) + f_{c\_cl}(x) \\ {}_{g\_r}(x) &= f_{c\_sr}(x) + f_{c\_cr}(x) + f_{c\_sl}(x) + f_{c\_cl}(x) \end{aligned}\right\} . \qquad (4)$$

n the intact state, with single lateralized stimuli, only the ight-hand components' functions have absolute loadings n the global function orientate left, $f_{g\_l}(x)$, and both such loadings are equal (i.e. each has a loading on that function of +0.5). Corresponding loadings exist for the left structures on global function orientate right, $f_{g\_r}(x)$. Because of the balanced, functioning systems yielded by the cortical and subcortical lesion pairs (i.e. the classical Sprague paradox cases), the share in the global functions of the collicular and cortical stations must be about equal, and the loadings must be 0.5 each. Were it otherwise, some degree of imbalance, manifest in an associated degree of impaired contralateral orienting would be evident in these cases.

The relationships between the component functions and sensory input are specified by the following equations, that embody the competitive nature of the interaction between the left and right sides:

$$\left.\begin{aligned} f_{c\_sr}(x) &= \begin{cases} +1 \text{ when } x = r \\ 0 \text{ when } x \neq r \end{cases} \\ f_{c\_cr}(x) &= \begin{cases} +1 \text{ when } x = r \\ 0 \text{ when } x \neq r \end{cases} \\ f_{c\_sl}(x) &= \begin{cases} 0 \text{ when } x \neq l \\ -1 \text{ when } x = l \end{cases} \\ f_{c\_cl}(x) &= \begin{cases} 0 \text{ when } x \neq l \\ -1 \text{ when } x = l \end{cases} \end{aligned}\right\} . \qquad (5)$$

We now reformulate the global functions, $f_{g\_l}(x)$ and $f_{g\_r}(x)$, in terms of the sum of the component functions, $f_c(x)$,

$$\left.\begin{aligned} f_{g\_r}(x) &= \begin{cases} r \text{ when } \sum f_c > 0 \\ null \text{ when } \sum f_c \leqslant 0 \end{cases} \\ f_{g\_l}(x) &= \begin{cases} null \text{ when } \sum f_c \geqslant 0 \\ l \text{ when } \sum f_c < 0 \end{cases} \end{aligned}\right\} . \qquad (6)$$

What lesions would be required to recover the loadings of the component functions, $f_c(x)$, on each of the global functions, $f_{g\_r}(x)$ and $f_{g\_l}(x)$, in this ideally simple, though empirically motivated, situation?

First, any single lesion will abolish contralateral orientating, because it yields an unbalanced system that is not captured appropriately by stimulus-related activity. Consider, for example, a lesion in the right superior colliculus abolishing the component function $f_{c\_sr}(x)$. If we present a stimulus $x=l$, on the left, then the sum of the outputs of the component functions $\sum f_c(x) = 0$ (equation (5)). Therefore, the animal will not orient either left or right (equation (6)) and the global function $f_{g\_l}(x)$ has been abolished. However, a stimulus presented on the right will still produce the correct orientating response, as the sum of the outputs of the component functions $\sum f_c(x) = -2$ (equation (6)). This would lead to the wrong conclusion, that the component function $f_{c\_sr}(x)$ had a weighting of 1 on the global function $f_{g\_l}(x)$. Lesioning a single structure is therefore insufficient, even in this very simple system, echoing from a different perspective the inferential inadequacy of single dissociations.

Second, double dissociations of $f_{g\_l}(x)$ and $f_{g\_r}(x)$ formed by pairs of independent single lesions of contraıteral structures, inherit precisely the same incorrect ttribution as was made for the single lesions and single issociations that comprise each double dissociation. Each onstituent single dissociation still incorrectly suggests a ɔading of 1.0 for any single structure. Hence, double issociations do not provide any further basis for recoering the loadings, echoing from this different perspecıve their inferential inadequacy.

Third, pairs of lesions will have variable effects, epending on whether the lesions are ipsi- or contraıteral. Pairs of ipsilateral lesions will abolish orientating ɔ the contralateral hemisphere, while contralateral ɛsions will produce paradoxical restoration of function ɛquations (5) and (6)). Therefore, ipsilateral paired ɛsions suggest a summed loading of one on the global ınction for both the lesioned structures (correct). Contralateral lesions suggest a summed loading of zero of ıe lesioned structures on the global function (incorrect). łowever, the cases in which contralateral pairs involve ollicular and cortical lesions show intact orientating ınctions. These cases reveal that the colliculus and ortex contribute equally to each $f_g(x)$, and so must have qual loadings on each global function. Because there are nly two structures on each side, this indicates that the ɔading of each structure's function on the contralateral lobal function must be 0.5 (correct).

Fourth, any odd number of lesions will always yield the bolition of orientating to the side contralateral to the ırger number of lesions. For example, simultaneous ɛsions to the left cortex and colliculus and the right ortex will abolish component functions $f_{c\_cr}(x)$, $f_{c\_cl}(x)$ nd $f_{c\_sl}(x)$, leaving only $f_{c\_cl}(x)$. This remaining right olliculus will allow orientating to the left (equations (5) nd (6)). This will suggest a loading of one on the global ınction $f_{g\_l}(x)$ for the single remaining colliculus (incorɛct). Hence, just as for single lesions, odd numbers of ɛsions do not allow the recovery of the true loadings in ıe intact system.

Fifth, quadruple lesions will abolish both global funcıons in this simple network (it remains to be seen what ıis pattern of inactivation will yield in the real brain), roviding no further help in recovering the precise indiviual loadings, but will correctly identify the summed ɔadings of all the structures.

Hence, in this minimal system there are lesion combiıations that can recover the loadings precisely, and so npute function to structure reliably. In this case, neither ngle nor double dissociations provided the necessary ıformation, but the paradoxically restored cases, partiularly those involving lesions in structures that were not ilateral mirrors of one another, allowed recovery of the ɔadings. However, the analysis above represents a ecomposition of the system close to being complete. The aradoxical restoration cases alone would not have rovided enough information to recover the loadings ʻithout knowledge of the connectivity of the system, and ʻithout knowledge of the importance of balanced competition in this system, which latter was derived in part rom the effects of the other lesion combinations. On the ne hand, then, these results suggest the optimistic onclusion that there are circumstances in which ınctions can be imputed to structures reliably. On the ther hand, a near-complete decomposition of this simple network was necessary to impute functions to its structures. This suggests that the problem of imputing function to structure from lesion effects may not be tractable by these means alone in the real brain, where a complete decomposition cannot be envisaged. It may, however, be possible to use other information about the organization of the network to reduce the necessity for exhaustive search. Structures and systems likely to possess negligible loadings on the global functions being tested could be excluded on the basis of membership of different connectional groupings (e.g. Burns and Young, this issue; Hilgetag *et al.* 1999), or by reference to activation during testing (see §6).

6. DISCUSSION

Very many insights into which brain component does what have been derived from examining what people or other animals do less well when particular brain structures are damaged. Whether this information is reliable, and whether reliable information can be gathered in future from this approach, are important issues. To address these issues, we have attempted to derive elements of a relationship between this process of imputation of functions to structures and the connectivity that we assume determines in part the effects of localized lesions. Through simulating the effects of lesioning stations in the thalamocortical network of the cat (§2), we determined three likely features of interactions between brain structures after a lesion. The consequences of these effects for the conventional patterns of inference in single dissociation emphasized the concerns from empirical studies that such inferences will sometimes be invalid. In addition, the consequences of the lesion effects, in common with results from empirical studies, suggested that double dissociation is no more reliable a means of imputing function to structure than single dissociation.

The characteristics of electronic circuits, and the limitations of what can be determined about the roles of their components, have been described by electrical engineers (e.g. Lewis 1970). For circuits with properties like those presumed for the brain, such as the importance of the context of the rest of the network, the prognosis for determining the roles of individual components from alterations of the behaviour of the system is extremely poor (e.g. Lewis 1970). In the most likely case, complete decomposition would be required. Buoyed, however, by the fact that imputations of function in the brain derived from lesion experiments have often been supported by other methods, we attempted to clarify the concepts of function and explored a more formal approach for imputing function to structure on the basis of the effects of brain lesions (§4). We found in §5, through a worked example of this approach, that it was possible to recover detailed and reliable information on the importance of particular structures to particular functions. Unfortunately, though, a comprehensive decomposition of our simple network appeared necessary to accomplish this. Because the large number of lesion experiments required to take the same approach to the brain cannot be envisaged, the prospects for deriving reliable imputations of function to structure in the brain by these means do not appear great.

One conclusion, then, is that our results suggest that all resently conceived rules of inference, both conventional nd the more formal approach we have developed above, re inadequate to impute functions to brain structures on he basis of lesion effects. This is as predicted from ystems theory (e.g. Lewis 1970). Another conclusion, owever, is that the propagated effects of lesions, the easons for the failure of conventional inferences and our nore formal approach suggest a possible way forward. Reliable inference appears to require exhaustive search hrough lesions of every station. Meeting this requirement ; plainly impractical in the brain. Multiple sources of nformation, though, could be brought to bear on two key ssues. Information from other methodologies might first e used to exclude many structures from the required earch, on the grounds that their loadings on the ehavioural function are likely to be negligible. Decomosition by inactivation could then be brought within ractical bounds. Information from other methods might lso be used in conjunction with inactivations to deternine the direct and indirect effects of the inactivations on ther stations. We note in this context that reverse engieering, for example of a faulty amplifier made elsewhere, ; typically carried out by reference to more information han the changes in input–output characteristics on emoval of internal components. In general, a known ignal is introduced, and a combination of electrical earch for the propagation of the signal through the ircuits, removal of components and observation of the utput is undertaken. A circuit diagram that describes the onnectivity and organization of the amplifier's subystems is often very helpful, mainly through excluding vhole regions of the system from consideration when aults are of a particular kind.

An analogous strategy could be implemented in the rain. Successors to the framework we developed in § 4 ould be used to specify the problem of identifying the oles of brain processors in some behavioural function. nformation on connectivity, such as indications of trongly intra-connected clusters of areas (e.g. Hilgetag, Burns, O'Neill, Scannell & Young, this issue; Young *et al.* 995; Burns & Young, this issue), could be used in onjunction with physiological information to identify kely stations and systems of interest, and systems unliely to be strongly involved in the function. Imaging pproaches could perhaps be employed to further deternine or cross-validate those stations and systems less nvolved in mediating a particular function, although not ll the links between imaging signals, blood, metabolism, euronal population dynamics and functional informaion processing changes are established, and some seem ot to be straightforward (Scannell & Young 1999). Patterns of inactivation effects, particularly in combinaion with concurrent information on activity, could then e interpreted rigorously in the context of an analytical ramework. In this framework, knowledge of the connecivity is a necessary but insufficient condition for reliable nference, which in this case would be constrained by nultiple, interacting sources of experimental information. n this way, a bridge between connectivity and the effects n behavioural function of lesions might be used to emonstrate principles and test concepts about a wide ariety of structure–function relationships and suggest further experiments using a wide variety of neuroscience methodologies.

REFERENCES

Barton, R. A. 1998 Visual specialization and brain evolution in primates. *Proc. R. Soc. Lond.* B **265**,1933–1937.

Barton, R. A. & Dean, P. 1993 Comparative evidence indicating neural specialization for predatory behaviour in mammals. *Proc. R. Soc. Lond.* B **254**, 63–68.

Colby, C. L. & Duhamel, J.-R. 1991 Heterogeneity of extrastriate visual areas and mulipl t parietal areas in the macaque monkey. *Neuropsychologia* **29**, 517–537.

Cosmides, L. & Tooby, J. 1995 From function to structure: the role of evolutionary biology and computational theories in cognitive neuroscience. In *The cognitive neurosciences* (ed. M. S. Gazzaniga), pp. 1139–1210. Cambridge, MA: MIT Press.

Damasio, H. & Damasio, A. R. 1989 *Lesion analysis in neuropsychology.* New York: Oxford University Press.

Dean, P. 1982 Analysis of visual behaviour in monkeys with inferotemporal lesions. In *Analysis of visual behaviour* (ed. D. Ingle, M. Goodale & R. Mansfield), pp. 587–618. Cambridge, MA: MIT Press.

Douglas, R. J. & Martin, K. A. C. 1991 Opening the grey box. *Trends Neurosci.* **14**, 286–293.

Douglas, R. J. & Martin, K. A. C. 1994 The canonical microcircuit: a co-operative neuronal network for neocortex. In *Structural and functional organisation of the neocortex* (ed. B. Albowitz, K. Albus, U. Kuhnt, H.-C. Nothdurft & P. Wahle), pp. 131–141. Berlin: Springer.

Douglas, R. J., Mahowald, M., Martin, K. A. C. & Stratford, K. J. 1996 The role of synapses in cortical computation. *J. Neurocytol.* **25**, 893–911.

Ennaceur, A., Neave, N. & Aggleton, J. P. 1997 Spontaneous object recognition and object location memory in rats: the effects of lesions in the cingulate cortices, the medial prefrontal cortex, the cingulum bundle and the fornix. *Exp. Brain Res.* **113**, 509–519.

Flechsig, P. E. 1905 Gehirnphysiologie und Willestheorien. In *Proceedings of the Fifth International Psychological Congress*, pp. 73–89. Translated by G. von Bonin 1960 In *Some papers on the cerebral cortex.* Springfield, IL: C. C. Thomas.

Grobstein, P. 1990. Strategies for analysing complex organization in the nervous system. I. Lesion experiments. In *Computational neuroscience* (ed. E. Schwartz). Cambridge, MA, and London: MIT Press.

Heywood, C. A., Gaffan, D. & Cowey, A. 1995 Cerebral achromatopsia in monkeys. *Eur. J. Neurosci.* **7**, 1064–1073.

Hilgetag, C.-C., O'Neill, M. A. & Young, M. P. 1996 Indeterminate organization of the visual hierarchy. *Science* **271**, 776–777.

Hilgetag, C.-C., Kötter, R. & Young, M. P. 1999 Paradoxical restoration of function: a mathematical model based on anatomical connectivity. *Prog. Brain Res.* **121**, 121–141.

Hunt, P. R. & Aggleton, J. P. 1998 Neurotoxic lesions of the dorsomedial thalamus impair the acquisition but not the performance of delayed matching to place by rats: a deficit in shifting response rules. *J. Neurosci.* **18**, 10 045–10 052.

Jones, G. V. 1983 On double dissociation of function. *Neuropsychologia* **21**, 397–400.

Killcross, S., Robbins, T. W. & Everitt, B. J. 1997 Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. *Nature* **388**, 377–380.

Krebs, J. R. & Davies, N. B. 1978 *Behavioural ecology, an evolutionary approach*, 1st edn. Oxford, UK: Blackwell Science.

Krebs, J. R. & Davies, N. B. 1991 *Behavioural ecology, an evolutionary approach*, 4th edn. Oxford, UK: Blackwell Science.

ashley, K. S. 1952 Functional interpretation of anatomic patterns. *Res. Pub. Assoc. Nervous Mental Dis.* **30**, 529–547.

ewis, E. R. 1970 Neural subsystems: goals, concepts, and tools. In *The neurosciences second study program* (ed. F. O. Schmitt), pp. 384–396. New York: Rockefeller University Press.

omber, S. G. & Payne, B. R. 1996 Removal of 2 halves restores the whole—reversal of visual hemineglect during bilateral cortical or collicular inactivation in the cat. *Vis. Neurosci.* **13**, 1143–1156.

uria, A. R. 1973 *The working brain.* New York: Penguin Books.

Ierabet, L., Desautels, A., Minville, K. & Casanova C. 1998 Motion integration in a thalamic visual nucleus. *Nature* **396**, 265–268.

Ieynert, T. 1890 Uber das Zusammenwirken der Gehirntheile. Verhandlungen des 10th Internat. Mediz. Kongress, Berlin, vol. 1, pp. 173–190. Translated by G. von Bonin 1960 In *Some papers on the cerebral cortex.* Springfield, IL: C. C. Thomas.

Ionakow, C. 1910 *Uber Lokalisation der Hirnfunktionen.* Wiesbaden.

Ionakow, C. 1914 *Die Lokalisation im Grosshirn und der Abbau der Funktionen durch corticale Herde.* Wiesbaden: Bergmann.

sorio, D. & Vorobyev, M. 1996 Colour vision as an adaptation to frugivory in primates. *Proc. R. Soc. Lond.* B **263**, 593–599.

ayne, B. R., Lomber, S. G., Geeraerts, S., Vandergucht, E. & Vandenbussche, E. 1996*a* Reversible visual hemineglect. *Proc. Natl Acad. Sci. USA* **93**, 290–294.

ayne, B. R., Lomber, S. G., Villa, A. E. & Bullier, J. 1996*b* Reversible deactivation of cerebral network components. *Trends Neurosci.* **19**, 535–542.

agozzino, M. E., Adams, S. & Kesner, R. P. 1998 Differential involvement of the dorsal anterior cingulate and prelimbic-infralimbic areas of the rodent prefrontal cortex in spatial working memory. *Behav. Neurosci.* **112**, 293–303.

ahakian, B. J., Semple, J., Polkey, C. E. & Robbins, T. W. 1995 Visuospatial short-term recognition memory and learning after temporal-lobe excisions, frontal-lobe excisions or amygdalo-hippocampectomy in Man. *Neuropsychologia* **33**, 1–24.

Scannell, J. W., Blakemore, C. & Young, M. P. 1996 Analysis of connectivity in the cat cerebral cortex. *J. Neurosci* **15**, 1463–1483.

Scannell, J. W., Burns, G., O'Neill, M. A. & Young, M. P. 1997 The organisation of the thalamocortical network of the cat. *Soc. Neurosci. Abstr.* 23.

Scannell, J. W., Burns, G., O'Neill, M. A. & Young, M. P. 1999 The connectional organisation of the thalamo-cortico-cortical system of the cat. *Cerebr. Cortex* **9**, 277–299.

Scannell, J. W. & Young M. P. 1999 Population activity and functional imaging. *Proc. R. Soc. Lond.* B **266**, 875–881.

Selden, N. R. W., Everitt, B. J., Jarrard, L. E. & Robbins, T. W. 1991 Complementary roles for the amygdala and hippocampus in aversive-conditioning to explicit and contextual cues. *Neuroscience* **42**, 335–350.

Shallice, T. 1988 *From neuropsychology to mental structure.* Cambridge University Press.

Sprague, J. M. 1966 Interaction of cortex and superior colliculus in mediation of visually guided behavior in the cat. *Science* **153**, 1544–1547.

Teuber, H.-L. 1955 Physiological psychology. *A. Rev. Psychol.* **6**, 267–296.

Turner, D. C. & Bateson, P. P. G. 1986 *The domestic cat: the biology of its behaviour.* Cambridge University Press.

Wallace, S. F., Rosenquist, A. C. & Sprague, J. M. 1989 Recovery from cortical blindness mediated by destruction of nontectotectal fibers in the commissure of the superior colliculus in the cat. *J. Comp. Neurol.* **284**, 429–450.

Wallace, S. F., Rosenquist, A. C. & Sprague, J. M. 1990 Ibotenic acid lesions of the lateral substantia nigra restore visual orientation behavior in the hemianopic cat. *J. Comp. Neurol.* **296**, 222–252.

Young, M. P., Scannell, J. W., O'Neill, M. A., Hilgetag, C. C., Burns, G. & Blakemore, C. 1995 Non-metric multidimensional scaling in the analysis of neuroanatomical connection data and the organization of the primate cortical visual system. *Phil. Trans. R. Soc. Lond.* B **348**, 281–308.