Degeneracy and cognitive anatomy

Cathy J. Price and Karl J. Friston

Cognitive models indicate that there are multiple ways of completing the same task. This implicit degeneracy cannot be revealed by functional imaging studies of normal subjects if more than one of the sufficient neural systems is activated. Nor can it be detected by neuropsychological studies of patients because their performance might not be impaired when only one degenerate system is damaged. We propose that degenerate sets of sufficient neural systems can only be identified by an iterative approach that integrates the lesion-deficit model and functional imaging studies of normal and neurologically damaged subjects.

The term 'degeneracy' was introduced to neurobiology by Edelman and colleagues [1–4], who defined it as 'the ability of elements that are structurally different to perform the same function or yield the same output.' Edelman and Gally [4] provide examples of degeneracy at many different levels of biological organization, ranging from the genetic code (different nucleotide sequences encode the same polypeptide) to communication (equivalent but non-identical structures convey the same meaning). The structural elements range from molecular, through neuronal and cortical levels through to non-physical structures such as computational modules in cognitive science. This article is concerned with a structural level that is appropriate for neuropsychological and neuroimaging enquiry: namely, cortical and subcortical regions of the brain with a spatial scale of millimetres to centimetres. These regions can be assembled into neuronal systems that are sufficient for the cognitive operations required for task performance. Figure 1 illustrates the elaboration of different neuronal systems. The simplest – functional segregation – ascribes a particular cognitive function or processing capacity to a single area that is both necessary and sufficient. The functional integration perspective allows for a more distributed architecture where interactions between two or more brain areas (a system) are necessary. Functional degeneracy implies that there are degenerate sets of areas in which each system is sufficient. The only necessary brain areas occupy the intersection of all sufficient systems. This might involve a subset, or none, of the areas that comprise sufficient systems. A useful measure of the degree of degeneracy is afforded by its order. The order of degeneracy refers to the number of disjoint sufficient systems and, at an operational level, it can be defined by the minimum number of areas that must be removed before function is lost (i.e. one area for first-order degeneracy, two areas for second-order degeneracy).

Cathy J. Price
Wellcome Dept of Imaging Neuroscience, Institute of Neurology, Queen Square, London, UK WC1N 3BG.
e-mail: cprice@fil.ion.ucl.ac.uk

The ability of structurally different mechanisms to yield the same output is well appreciated in cognitive neuroscience. For example, neuropsychological data indicate that familiar, regularly spelled words can be read via either spelling–sound relationships or lexical/semantic processes [5,6]. Regular word reading is, therefore, left relatively intact following damage to only one cognitive mechanism. Likewise, neuropsychological studies show that objects can be recognized either on the basis of their global shape or by the presence of distinguishing features [7]. In other words, different cognitive functions (global and local feature processing) yield the same output (object recognition); see Fig. 2 for further details. In the neuropsychological literature, different mechanisms for the same task are usually referred to in terms of cognitive strategies. We use the term degeneracy because it characterizes the structure–function relationship and indicates that there is more than one neuronal system for producing the same response. The characterization of cognitive anatomy then reduces to identifying these systems.

Differentiating degenerate sets of neuronal systems for the same task might provide important clues to how sensorimotor and cognitive functions recover following neurological damage. This is because degeneracy clearly underlies recovery by providing robustness to failure or damage. When degenerate sets of neuronal systems are available, damage to one system does not impair response accuracy, which is retained by virtue of the remaining systems. Response times might be affected but might also recover rapidly following compensatory adjustments within the remaining systems. Degeneracy also enables new learning because previous learning, which is embodied in the other systems, is not lost following plastic changes to any single system. This general robustness to either local damage or new learning, is closely related to 'graceful degradation' in the context of parallel distributed systems [8,9]. Furthermore, on an evolutionary scale, it is clear that degeneracy is essential because selective mechanisms can only act if there is more than one phenotypic mechanism to select from [4].

Despite the importance of degeneracy, alternative neuronal systems for the same function cannot be revealed by lesion or functional imaging studies conducted in isolation. We review the limitations that degeneracy imposes on these techniques. We then propose that degenerate sets of sufficient neural systems can only be identified by an iterative approach that integrates data from lesion studies with functional imaging of normal and neurologically damaged subjects.

The lesion-deficit model

Until the first PET studies of language processing were reported [10], the primary method for establishing cognitive anatomy was the lesion-deficit model. The principle behind this approach is that if a

 1364-6613/02/$ – see front matter PII: S1364-6613(02)01976-9

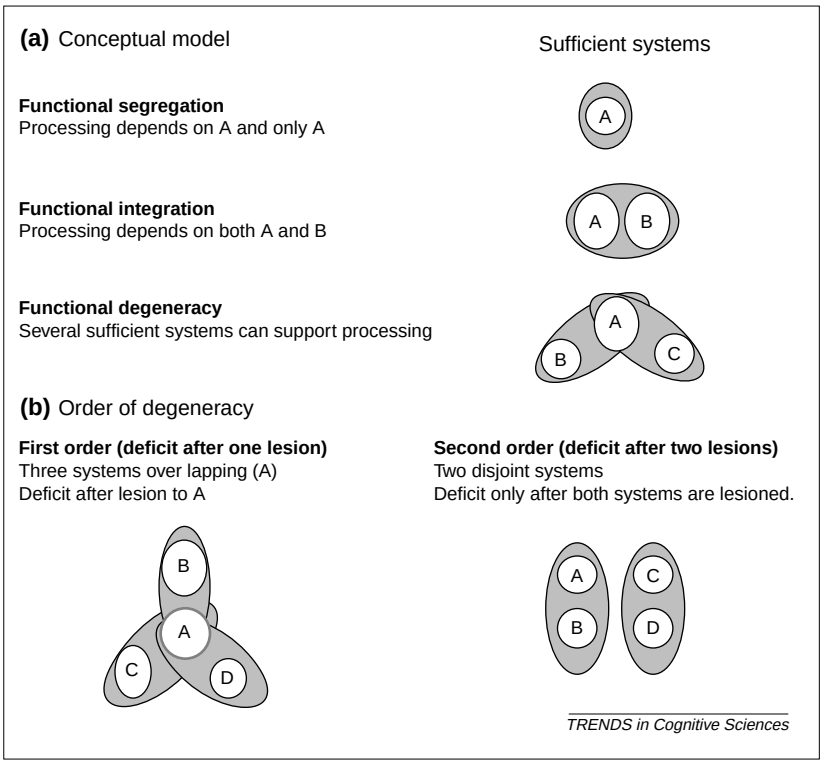

Fig. 1. Segregation, integration and degeneracy. (a) Functional brain architectures. White circles denote brain areas and grey ellipses encompass sets of regions that are sufficient to complete the task. (b) The distinction between disjoint and overlapping sufficient systems. Unlike overlapping systems (left), in the disjoint organization (right), no single area is necessary and sufficient. A useful measure of the degree of degeneracy is afforded by its order, which is defined by the minimum number of areas that must be removed before function is lost (i.e. one for first-order degeneracy, two for second-order degeneracy). Note that first-order degeneracy is not the same as the absence of degeneracy. For example, a single lesion will induce a deficit in all three functional architectures in (a). However, only the last example has more than one sufficient system and shows first-order degeneracy.

physiological lesion results in a cognitive deficit, then part of the lesioned area must be necessary for the lost cognitive process. Throughout the last century several methodological developments enabled the lesion-deficit approach in humans. Delineation of permanent (pathophysiological and traumatic) lesions was initially dependent on postmortem studies until the 1970s–1980s when *in vivo*, 3-D structural-imaging techniques, such as computerized tomography and magnetic resonance imaging, became available. Lesions can also be induced transiently either electrically [11–13], pharmacologically [14] or using transcranial magnetic stimulation (TMS). In particular, TMS provides the first means to systematically and non-invasively investigate the effect of a variety of different lesions in normal brains [15,16].

Irrespective of the technique used, the lesion-deficit model has well recognized limitations. For example, the full set of regions that comprise a neural system is difficult to establish because some areas are resistant to either ischemic damage (e.g. areas that have more than one blood supply) or TMS, which cannot reach deep structures. The primary limitation we focus on, however, which was documented in the early part of the 20th century, relates to observations that similar lesions can result in very different effects in different subjects and the effect of the same lesion can vary over time in a subject (e.g. over the course of recovery) [17–20]. To explain these inconsistencies, Lashley [19,20] proposed two controversial theories. The first was the theory of 'mass action', which stated that performance depends on the total available cortex after damage (rather than which areas were damaged). The second was the theory of 'equi-potentiality', which stated that an area can have different functions depending on need. These theories emphasize distributed rather than localized processing and explain the variable effects of different lesions. However, they are incompatible with what we know about modularity and anatomical segregation of sensorimotor and cognitive functions [21]. The concept of degeneracy, by contrast, accommodates intersubject variability as well as functional specialization. Whereas equi-potentiality implies that any brain system can take over a function, the concept of degenerate brain systems proposes that there might be a limited number of specialized systems for the same function.

The crucial implication of degeneracy for the lesion-deficit model is that when there is localized damage to a degenerate system, there may be no cognitive deficit, and when there is no cognitive deficit, we cannot deduce whether the damaged area is part of a system that would have been sufficient to support the cognitive process. In other words, degenerate brain systems cannot be identified using the lesion-deficit model because, when one system can substitute for another, the ability to complete the task will be protected from selective damage to any one system. There might, therefore, be many cases of degeneracy in the normal brain that have not been detected during routine neuropsychological investigations because patients continue to provide correct responses when only one system is damaged.

Functional imaging studies of normal subjects

It is generally assumed that functional imaging techniques will overcome many of the limitations associated with the lesion-deficit model. The main advantages that functional imaging offers are that it can identify the set of regions that are engaged for one task relative to another and it is not limited to 'lesioned' cortex. For example, when normal subjects are instructed to match written words on the basis of their meaning, they activate a distributed set of areas (the semantic-retrieval network) relative to when they match the same words on the basis of stimulus size (Fig. 3a) [22,23]. The set of activated areas can, therefore, be considered sufficient for the cognitive operations that differentiate the two tasks (this ignores sensitivity issues and assumes that we can detect activation if the brain responds to a task manipulation). Activations might also highlight regions that were not previously considered important for semantic retrieval on the basis of lesion

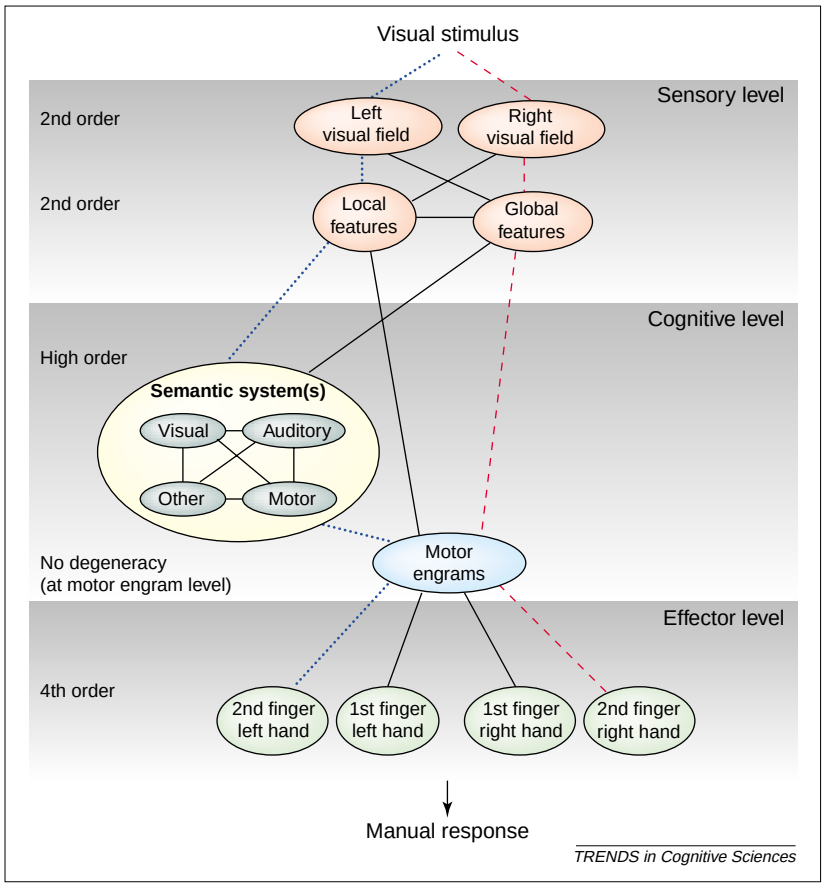

Fig. 2. Order and levels of degeneracy. The sensorimotor and cognitive components of a simple stimulus–response task. At a sensory level, visual stimuli can be processed by either the left or right visual fields and require either global or local feature processing. The cognitive level distinguishes between semantic and non-semantic routes to motor engrams and, then, at the effector level, any of, say, four fingers can make the required manual response. The task design/analysis and the nature of the psychological or neurophysiological measurements implicitly define the level of degeneracy being investigated. The order of degeneracy, defined by the number of disjoint systems that can perform the same operation, can change with either the structural level or the function. For example, there might be no degeneracy in cortical regions but high-order degeneracy at the neuronal level (if many neurons or neuronal assemblies have to be eliminated before function is compromised). Similarly, the order of degeneracy is sensitive to the complexity of the function. For example, if the structural elements are fingers and thumbs, and any one of 10 fingers can be used to press a button, the order for a single button-press would be 10. For piano playing all 10 fingers/thumbs are required (i.e. there is low or no degeneracy).

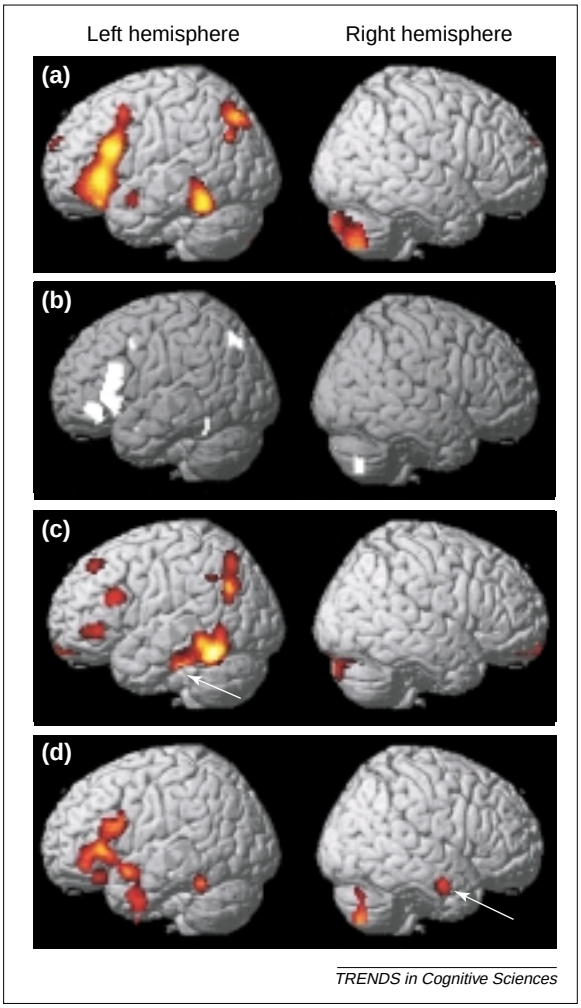

Fig. 3. Normal activation patterns during a semantic paradigm. (a) The results of a fixed effect analysis ($p<0.05$ corrected for multiple comparisons) averaged over 12 different subjects. This analysis does not distinguish effects that are common to all subjects from those that are evoked by a subset of subjects. By contrast, (b) depicts only those voxels that were activated in all participating subjects (conjunction of activation in each subject) and illustrates consistency across subjects in a common set of cortical areas. Data from subject 1 only (c) and subject 2 only (d) reveal inter-subject variation (white arrows) that is hidden in (a) and (b). These analyses based on subject-specific effects indicate areas that might be components of different systems. The methods and details have been reported previously [27,28].

studies alone. However, functional imaging cannot determine whether the activated areas are necessary for task performance. For example, if two or more degenerate systems are capable of providing correct responses on a semantic task, none of the individual activated areas might be necessary. Furthermore, degenerate systems might activate in parallel or only one system might be engaged at a time. Crucially, functional imaging studies of normal subjects cannot determine how many systems are jointly activated or reveal the systems that are not activated. In addition, if different subjects engage different systems, then there might be no common pattern of activation and intersubject variation will be high. This is illustrated in Fig. 3, which shows a contrasting pattern of activation in Subjects 1 and 2. Whereas Subject 1 activated an anterior region of the left inferior temporal cortex, Subject 2 activated a right anterior middle temporal area that was not common to all subjects. Functional imaging studies normally discard these 'idiosyncratic' activations and treat them as random error. Nevertheless, we cannot exclude the possibility that they are not random but reflect degenerate mechanisms for performing the same function.

Combining the lesion-deficit approach with functional imaging studies of normal subjects

Functional imaging can, in principle, identify the set of regions that are sufficient for a cognitive operation and the lesion-deficit model identifies which of these areas are necessary. We have previously argued [24,25] that a combination of functional imaging and the lesion-deficit model should enable us to identify

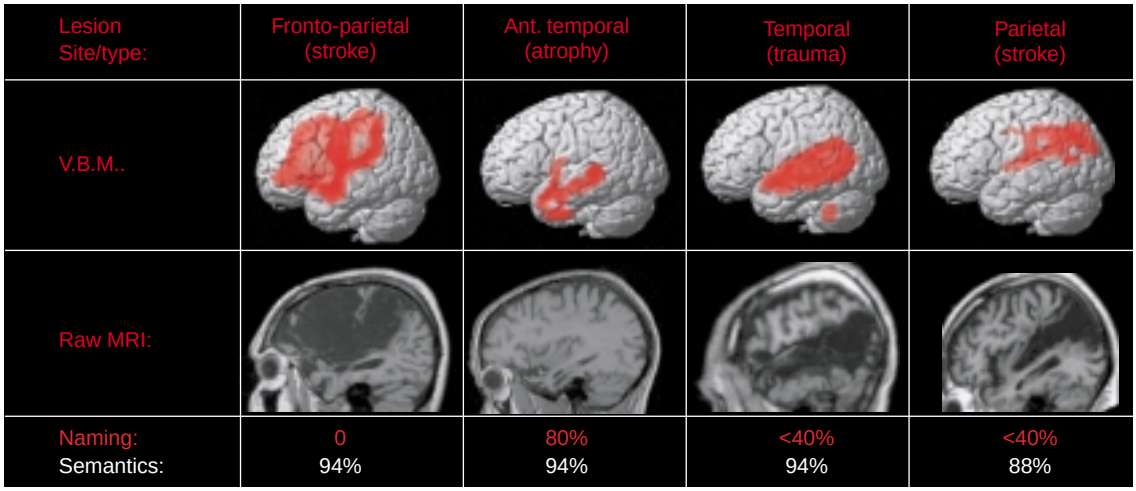

Fig. 4. Combining functional imaging and the lesion-deficit model. The site, type and extent of lesions (red, identified from voxel-based morphometry, VBM [29]) in areas that were activated in normal subjects (see Fig. 3a). Percentages refer to the patients' accuracy on picture naming and the 'Pyramids and Palm Trees' (semantic) task [30]. Thus, none of the cortical areas activated in normal subjects appears necessary for correct performance on the semantic task.

'necessary and sufficient' brain systems. We were wrong – there might be no single necessary and sufficient system because the existence of two or more degenerate systems that do not overlap precludes the existence of a single necessary system or area. Figure 4 illustrates this point by showing that lesions encompassing each of the areas activated in our semantic paradigm do not dramatically impair performance accuracy. There must, therefore, be a high degree of degeneracy in the neural systems for this semantic task, with at least two disjoint systems that can provide correct responses.

If more than one system can perform the task, performance will only be affected when all possible systems are damaged. Thus, the components of degenerate systems can only be established by finding which combinations of lesions disrupt performance. For example, performance might be unaffected by a lesion to either the parietal area or the temporal area if they are components of different systems but if performance is impaired following lesions to both the parietal and temporal areas, we could deduce that either the temporal or parietal systems were necessary for performance. The obvious difficulty with this approach is that, when functional imaging indicates several candidate areas are involved in a cognitive task, many possible combinations of lesions need to be investigated before a performance deficit is observed. Nevertheless, specific hypotheses can be generated by examining intersubject variability in activation patterns. For example, the results in Figs 3c and 3d show that different subjects activate different brain areas, one engaged the right anterior temporal cortex (RATC) and one the left inferior temporal cortex (LITC). From these results we might hypothesize that there are different brain systems for the semantic task, one involving RATC and one involving LITC. We could then investigate how performance is affected by damage to (1) the RATC alone, (2) the LITC alone and (3) both the RATC and LITC. Indeed, Hodges and colleagues found that deficits on our semantic task are more pronounced following bilateral damage to the anterior temporal lobe [26] than damage to either the left or right only. These results indicate that RATC and LITC are parts of different systems, with one able to substitute for the other and neither necessary unless the other is damaged.

Although patients with multiple lesions to specific sites are rare, strokes frequently result in large lesions that encompass many brain regions. In the absence of imaging data, the lesion-deficit model is not usually particularly informative with such large lesions because multiple functions and brain areas are involved. However, for any given task integrating neuropsychological investigations and imaging results can recategorize extensive lesions in terms of whether there is damage to one, two, three, four or more regions activated in normal subjects.

Motivation for functional imaging studies of neurologically damaged patients

We have argued that functional imaging results can be used to guide neuropsychological investigations and that neuropsychological investigations can be used to test predictions generated from functional imaging data. It is possible, however, that the ability to complete the task might not be impaired following any combination of lesions to areas activated in functional imaging studies of normal subjects. This could occur if there are latent systems that are either untrained or 'inhibited' in normal individuals. Indeed, latent systems might only be activated if prepotent systems are lesioned. A complete perspective on degeneracy therefore necessitates functional imaging

Acknowledgements
This work was funded by the Wellcome Trust. The data reported in Figs 3, 4 and 5 were collected by C.J.P. and co-workers, and originally reported in a different format in Rik Vandenberghe *et al.* [22] and Cath Mummery *et al.* [23].

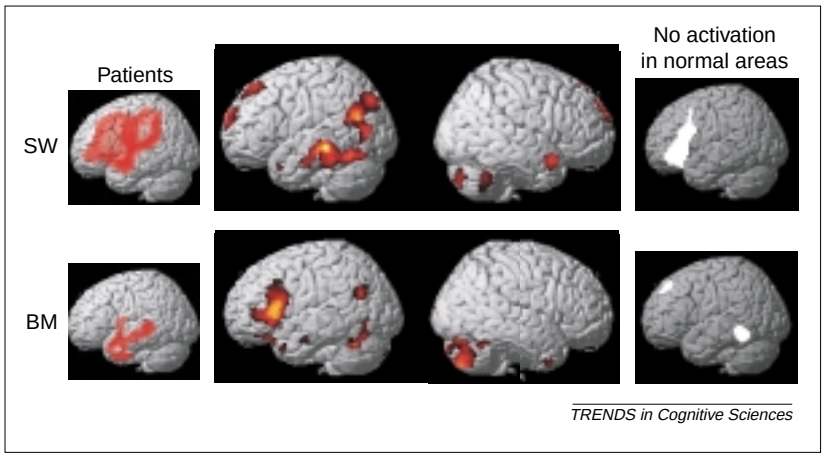

Fig. 5. Functional imaging in two neurologically damaged patients (SW and BM). Left to right: the lesions (red, identified by voxel-based morphometry), functional imaging results, and reduced activation (white) in areas activated in normal subjects on a semantic task (see Fig. 3). Both patients' performance accuracy was intact despite damage to components of the normal semantic retrieval system.

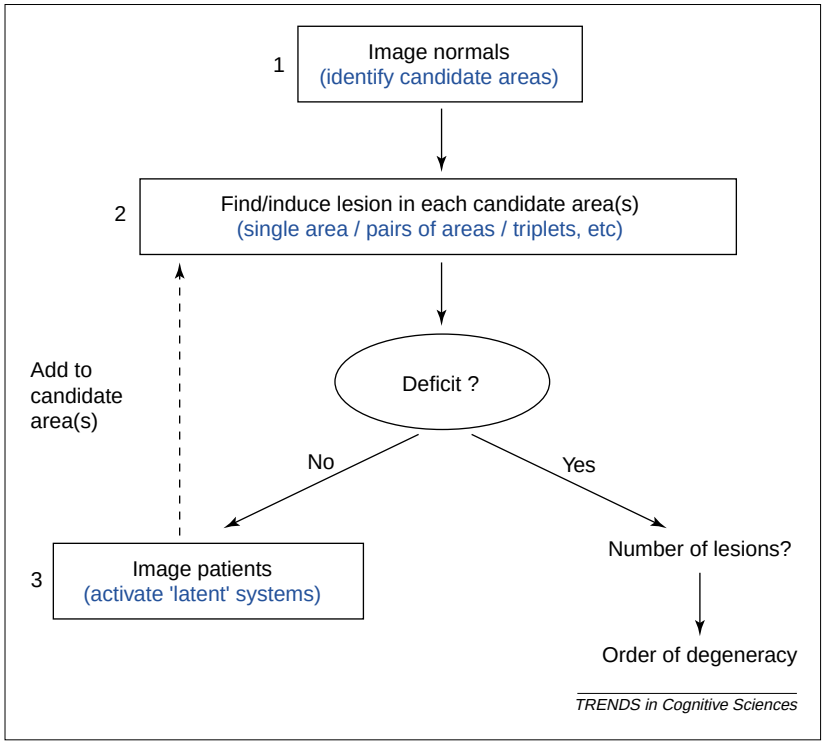

Fig. 6. A systematic approach to establish the order of degeneracy by combining neuroimaging and lesion data (neuropsychology or transcranial magnetic stimulation). First, functional imaging of normal subjects identifies candidate areas that are activated with one task component. Second, the effect of a 'lesion' (real or induced) in each of these areas is established. If a single lesion results in a deficit, the lesioned area must be a necessary component of all potential systems. If there is no deficit, higher-order degeneracy can be inferred and functional imaging of the lesioned brain is required to establish 'latent' areas (not activated in normal subjects). Once the set of candidate regions has been augmented, investigating the effect of lesions to pairs of regions can, in principle, reveal the order of degeneracy. This is repeated until the order of degeneracy is determined with a lesion that encompasses *n* regions.

studies of patients who can complete the task, despite damage to areas that are activated in normal subjects. In addition to revealing latent systems, functional imaging studies of neurologically damaged patients (1) increase intersubject variability, thereby allowing more specific hypotheses to be generated concerning the combinations of areas involved in different neural systems, (2) are required to exclude the possibility that the task can be completed because there is residual responsiveness in or around the lesion site, and (3) indicate how damage to one area affects other undamaged areas. For instance, damage to part of one system might result in underactivity (relative to normal) in the rest of that system and overactivity in the remaining intact systems. This would enable us to group areas together on the basis of how activation in undamaged areas is affected by the lesion.

Figure 5 illustrates the results of functional imaging experiments with two patients who provided accurate responses on a semantic task despite damage to components of the normal system [23,24]. By contrasting different patterns of activation, we can make predictions about the possible neural systems. Patient SW showed no inferior frontal activation but strong medial superior frontal activation. BM showed the reverse. The inferior and medial frontal areas therefore appear to be parts of different systems. We also predict that areas are more likely to be part of the same system if they coactivate in the same patient. Likewise, the undamaged areas with abnormal (reduced) responses are likely to be part of the same system as the damaged area. These predictions need to be tested with neuropsychological studies that investigate the effect of multiple lesions. Performance accuracy will be impaired if all possible systems are damaged but will not be impaired if damage only occurs to several components of the same system.

Conclusion

In this article we have discussed the concept of degeneracy at the level of cognitive anatomy. The key point is that a full understanding of the degenerate sets of neural systems that underlie any given cognitive task can only be revealed by systematically combining data from neuropsychological and neuroimaging studies of normal subjects and neurological patents (see Fig. 6). We believe that an understanding of degeneracy will provide insights into intersubject variability and the mechanisms that sustain recovery of cognitive function.

References

1 Edelman, G.M. (1978) *The Mindful Brain* (Edelman, G.M. and Mountcastle, V.B. eds.), pp. 51–100, MIT Press
2 Edelman, G.M. (1989) *The Remembered Present. A Biological Theory of Consciousness*, Basic Books
3 Tononi, G. *et al.* (1999) Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3257–3262
4 Edelman, G.M. and Gally, J.A. (2001) Deneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13763–13768
5 Marshall, J.C. and Newcombe, F. (1973) Patterns of paralexia: a psycholinguistic approach. *J. Psycholinguist. Res.* 2, 175–199
6 Seidenberg, M.S. and McClelland, J.L. (1989) A distributed developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568
7 Humphreys, G.W. and Riddoch, M.J. (1984) Routes to object constancy: implications from neurological impairments of object constancy. *Q. J. Exp. Psychol.* 36A, 385–415
8 Hinton, G. and Sejnowski, T. (1986) Learning and relearning in Boltzman machines. In *Parallel Distributed Processing.* (Rumelhart, D. and McClelland, J. eds.), pp. 282–317, MIT Press
9 McClelland, J. *et al.* (1995) Why there are complementary learning systems in the

hippocampus and neocortex–Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457
10 Petersen, S.E. *et al.* (1988) Positron emission tomographic studies of the cortical anatomy of single word processing. *Nature* 331, 585–589
11 Penfield, W. and Roberts, L. (1959) *Speech and Brain Mechanism*, Princeton University Press
12 Ojemann, G.A. (1979) Individual variability in cortical localization of language. *J. Neurosurg.* 50, 164–169
13 Nobre, A.C. *et al.* (1994) Word recognition in the human inferior temporal lobe. *Nature* 372, 260–263
14 Wada, J. and Rasmussen, T. (1960) Intracarotid injection of sodium amytal for the lateralization of speech dominance: experimental and clinical observations. *J. Neurosurg.* 17, 266–282
15 Pascual-Leone, A. *et al.* (1999) Transcranial magnetic stimulation: studying the brain-behaviour relationship by induction of 'virtual lesions'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1229–1238
16 Pascual-Leone, A. *et al.* (2000) Transcranial magnetic stimulation in cognitive neuroscience–virtual lesion, chronometry, and functional connectivity. *Curr. Opin. Neurobiol.* 10, 232–237
17 Marie, P. (1906a) Revision de la question de l'aphasie: La troisieme convolution frontale gauche ne joue aucun tole speciale dans la fonction du langage. *Semaine Medicale* 21, 241–247. (Repreinted in Cole, M.F. and Cole, M. eds., (1971), Pierre Marie's papers on speech disorders. New York: Hafner).
18 Marie, P. (1906b) Revision de la question de l'aphasie: Que faut-il penser des aphasies sous-corticales (aphasies pures)? *Semaine Medicale* 26, 493–500
19 Lashley, K.S. (1929) *Brain Mechanisms and Intelligence*, University of Chicago Press
20 Lashley, K.S. (1950) In search of the engram. In *Symposia for the Society for Experimental Biology, No, 4*, Cambridge University Press
21 Fodor, J.A. (1983) *The Modularity of Mind*, MIT Press
22 Vandenberghe, R. *et al.* (1996) Functional anatomy of a common semantic system for words and pictures. *Nature* 383, 254–256
23 Mummery, C.J. *et al.* (1999) Disrupted temporal lobe connections in semantic dementia. *Brain* 122, 61–73
24 Price, C.J. *et al.* (1999) Delineating necessary and sufficient neural systems with functional imaging studies of neuropsychological patients. *J. Cogn. Neurosci.* 11, 4371–4382
25 Price, C.J. and Friston, K.J. (1999) Scanning patients on tasks they can perform. *Hum. Brain Mapp.* 8, 102–108
26 Mummery, C.J. *et al.* (2000) A voxel based morphometry study of semantic dementia. The relation of temporal lobe atrophy to cognitive deficit. *Ann. Neurol.* 47, 36–45
27 Friston, K.J. *et al.* (1999) Multi-subject fMRI studies and conjunction analysis. *NeuroImage* 10, 385–396
28 Price, C.J. and Friston, K.J. (1997) Cognitive conjunctions: a new approach to brain activation experiments. *Neuroimage* 5, 261–270
29 Ashburner, J. and Friston, K.J. (2000) Voxel-based morphometry – the methods. *NeuroImage* 11, 805–821
30 Howard, D. and Patterson, K. (1992) *Pyramids and Palm Trees: A Test of Semantic Access from Pictures and Words*, Thames Valley, Bury St Edmunds

When a good fit can be bad

Mark A. Pitt and In Jae Myung

How should we select among computational models of cognition? Although it is commonplace to measure how well each model fits the data, this is insufficient. Good fits can be misleading because they can result from properties of the model that have nothing to do with it being a close approximation to the cognitive process of interest (e.g. overfitting). Selection methods are introduced that factor in these properties when measuring fit. Their success in outperforming standard goodness-of-fit measures stems from a focus on measuring the generalizability of a model's data-fitting abilities, which should be the goal of model selection.

Mark A. Pitt*
In Jae Myung
Dept of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210-1222, USA.
*e-mail: pitt.2@osu.edu

The explosion of interest in modeling cognitive processes over the past 20 years has fueled the cognitive sciences in many ways. Not only has it opened up new ways of thinking about research problems and possible solutions, but it has also enabled researchers to gain a better understanding of their theories by simulating a computational instantiation of it. Modeling is now sufficiently mainstream that one can get the impression that the models themselves are replacing the theories from which they evolved.

What has not kept pace with the advances and interest in modeling is the development of methods for evaluating and testing the models themselves. A model is not interchangeable with a theory, but only one of many possible quantitative representations of it. A thorough evaluation of a model requires methods that are sensitive to its quantitative form. Criteria used for evaluating theories [1], such as testing their performance in an experimental setting, do not speak to the quality of the choices that are made in building their quantitative counterparts (i.e. choice of parameters, how they are combined) or their ramifications. The paucity of such model selection methods is surprising given the centrality of the problem itself. What could be more fundamental than deciding between two alternative explanations of a cognitive process?

How *not* to compare models

Mathematical model are frequently tested against one another by evaluating how well each fits the data generated in an experiment or simulation. Such a test makes sense given that one criterion of model performance is that it reproduce the data. A goodness-of-fit measure (GOF; see Glossary) is invariably used to measure their adequacy in achieving this goal. What is measured is how much a model's predictions deviate from the observed data [2,3]. The model that provides the best fit (i.e. smallest deviation) is favored. The logic of this choice rests on the assumption that the model that provides the best fit to all data must be a closer approximation to the cognitive process under investigation than its competitors [4].

Such a conclusion is reasonable if measurements were made in a noise-free (i.e. errorless) system. One of the biggest challenges faced by cognitive scientists is that human and animal data are noisy. Error arises from several sources, such as the imprecision of our measurement tools, variation in participants and their performance over time. The problem of random

 1364-6613/02/$ – see front matter PII: S1364-6613(02)01964-2