LANGUAGE AND COGNITIVE PROCESSES, 1995, *10* (3/4), 227–264

Connectionist Modelling: Implications for Cognitive Neuropsychology

John A. Bullinaria and Nick Chater

Neural Networks Research Group, Department of Psychology, University of Edinburgh, Edinburgh, UK

We review here the logic of neuropsychological inference in the context of connectionist modelling, focusing on the inference from double dissociation to modularity of function. The results of an investigation into the effects of damage on a range of small artificial neural networks that have been trained to perform two distinct mappings (rules *vs* exceptions), suggest that a double dissociation is possible without modularity. However, when these studies are repeated using sufficiently larger and more distributed networks, which are presumably more psychologically and biologically relevant, double dissociations are not observed. Further analysis suggests that double dissociation between performance on rule-governed and exceptional items is only found when the contribution of individual units to the overall network performance is significant, and hence that such double dissociations are merely artefacts of scale. In large, fully distributed systems, a wide range of damage produces only a single dissociation in which the main regularities are selectively preserved. Thus, in this context, connectionism appears to create no additional problems for the traditional neuropsychological inference.

INTRODUCTION

Cognitive neuropsychology aims to inform theories of normal cognitive function by looking at the way in which the cognitive system breaks down in patients with brain damage. The inference from patterns of breakdown to normal function is, however, notoriously difficult, and the nature of such

Requests for reprints should be addressed to John A. Bullinaria, who is currently at the Centre for Speech and Language, Department of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, UK. Nick Chater is currently at the Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK.

This research was supported by Grant No. SPG 9029590 from the Joint Councils Initiative in Cognitive Science and HCI. The conjugate gradient training discussed in this paper was carried out using a modified version of the Xerion simulator originally developed by Tony Plate, Drew van Camp and Geoffrey Hinton. We would like to thank David Plaut, Gerry Altmann and an anonymous reviewer for their very useful comments on an earlier draft of the manuscript.

© 1995 Lawrence Erlbaum Associates Limited

inferences depends on the theories of normal function that are under consideration (Caramazza, 1984; Gregory, 1961; Shallice, 1988). The methodology of cognitive neuropsychology is rooted in "box and arrow" cognitive models, in which the architecture of the cognitive system is specified in very broad terms. Patterns of neuropsychological breakdown are assumed to correspond to selective damage to specific boxes and arrows. Conversely, observed patterns of deficit are used to constrain how such box and arrow models should look. The augmentation of the box and arrow models with neural network models of a wide range of the cognitive processes that neuropsychology has studied, thus poses the question: How, if at all, should the methodology of cognitive neuropsychology respond to the introduction of connectionist modelling techniques? It is this issue that this paper addresses.

We begin by considering the logic of cognitive neuropsychological inference in quite abstract terms, and then concentrate on a specific methodological principle, the inference from double dissociation to modularity of function. Double dissociation has been of central methodological importance because it promises to allow the neuropsychologist to map out the structure of the cognitive system. We review past work on the reliability of such inference for box and arrow models and in neural network models. We then present a range of simulations which show double dissociations between rule and sub-rule/exception performance in small feedforward neural networks. However, as we scale up towards larger, more realistic, distributed systems, only a single dissociation persists. The generality and implications of this work are considered and we suggest that some types of damage can be extrapolated more confidently than others from lesion studies on small-scale artificial neural network systems to patterns of breakdown that can be expected in the brain. Finally, we examine the methodological implications of neural network simulations for cognitive neuropsychology.

THE LOGIC OF NEUROPSYCHOLOGICAL INFERENCE

To elucidate the nature of practical neuropsychological inference, we first consider the ideal conditions for such inference, and then consider what simplifying assumptions must be made in practice, where such conditions do not generally hold.

In the ideal case, predictions concerning likely cognitive deficit can be derived if the cognitive system is understood (i) in terms of the computations being performed, (ii) the way that those computations are implemented in the neurophysiology of the brain, and (iii) if the specific damage that a particular patient or class of patients has suffered is known in detail[1] (for other discussions of the logic of neuropsychological inference, see Caramazza, 1986; Ellis, 1987; Shallice, 1988). Given these prerequisites, it is possible to derive predictions about the cognitive deficits that will be associated with each pattern of damage and these predictions can then be compared with observed cognitive deficits, and conjectures about (i), (ii) and (iii) can be revised accordingly. From the point of view of cognitive neuropsychology, interest focuses on how neuropsychological data can lead to the revision of (i), the computational theory of the cognitive system.

In practice, however, knowledge of (i), (ii) and (iii) is conjectural, and specified only in the broadest terms. Regarding (i), the cognitive system is often specified only at the level of large-scale architectural organisation, typically in the standard box and arrow notation. Recently, rather more detailed connectionist style models have also been considered. Regarding (ii), the neural implementation of cognitive processes is generally not explicitly considered at all, apart from some considerations of cerebral localisation, largely because detailed information is not available. Regarding (iii), the lesion damage can only be identified at a gross level, and damage is often diffuse in any case. Since (i), (ii) and (iii) are known in such little detail, detailed direct predictions of likely patterns of cognitive deficit cannot be derived and compared with the observed patterns of deficits found in neuropsychological patients. How, then, can neuropsychological data constrain cognitive theory?

A bold, but perilous, path is to make strong simplifying assumptions concerning (i)–(iii) in order to obtain predictions concerning likely patterns of damage. For box and arrow models, the key assumption is that brain damage leads to selective damage to particular "boxes" and "arrows". Furthermore, it is assumed that impaired performance reflects directly the operation of this damaged system, rather than being complicated by compensatory cognitive strategies (Plaut, in press a). A potential problem is that, even given this assumption, it may not be clear what predictions can be made, unless the boxes and arrows are specified in detail (Seidenberg, 1988). In neural network models, the crucial simplifying assumption is that brain damage can be modelled as involving the removal of, or disturbance to, particular processing units and/or connections. One attractive feature of such models is that, given this assumption, it is possible to derive detailed,

[1] There is a hidden assumption here, that aspects (i) and (ii) of the cognitive system are the same across the entire patient population. If different patients have different cognitive systems even in the normal state, attempts to infer these various structures from the range of observed pathologies will be incredibly difficult. The "principle of universality" (Caramazza, 1986) need not always be entirely an article of faith however, since evidence can be drawn from experimental studies on the normal population.

quantitative predictions (e.g. Bullinaria, 1994c; Hinton & Shallice, 1991; Patterson, Seidenberg, & McClelland, 1989; Plaut & Shallice, 1993; Plaut, in press a).

It is now clear how neuropsychological data can help decide between alternative cognitive level accounts. The predictions of each of these theories are derived, using appropriate simplifying assumptions. The degree to which the neuropsychological data favour one theory over the rest depends on (1) how well that theory predicts the data and (2) how well the other theories predict the data. That is, the degree to which neuropsychological data confirm a given theory depends on the extent to which that theory is relatively well able to predict those data when compared with alternative theories under consideration. A corollary of this view of the logic of neuropsychological inference is that the strength of the inference from data to a particular theory depends on which other theories *are* under consideration. In this paper, we shall be concerned with the degree to which extending the class of theories of cognitive processes to include connectionist theories requires revision of standard neuropsychological inferences, which were developed with a narrower class of theories—roughly box and arrow models—in mind.

In principle, any aspect of the behaviour observed in neuropsychological patients could be used as data to tell apart different cognitive theories. However, in practice, particular attention is paid to certain patterns of deficit across different patients, which are thought to provide especially strong evidence in deciding between theories. Specifically, we concentrate below on one pattern of patient data which has been viewed as central to neuropsychological inference to theories of normal function—that is, double dissociation.

THE DOUBLE DISSOCIATION INFERENCE

The cornerstone of cognitive neuropsychology is the inference from double dissociation (Teuber, 1955) to modularity of function. Two tasks, A and B, doubly dissociate across a patient population if there are some patients who have normal or near normal performance on A, but impaired performance on B, and there are some patients with the reverse pattern of deficit. The double dissociation inference passes from the observation of such a pattern of double dissociation to the conclusion that A and B cannot be subserved by the same cognitive machinery. More strictly, although tasks A and B may to some extent draw on the same aspects of the cognitive system, there must be parts of the cognitive system specific to A and others which are specific to B. In box and arrow terms, the conclusion is that at least some box or arrow must be involved in A but not in B, and some box or arrow must be involved in B but not in A.

The double dissociation inference has been applied across a broad range of cognitive tasks. An example that is particularly relevant to the discussion below is the dissociation between the reading of irregular words and the reading of nonwords: An extreme case of surface dyslexia, K.T. (McCarthy & Warrington, 1986), showed normal accuracy for reading nonwords and regular words, but was severely impaired at reading exception words. An extreme case of phonological dyslexia, W.B. (Funnell, 1983), could not read nonwords at all, but was still able to read most regular and irregular words (85% correct from a set of 712 widely varying words). This double dissociation between nonword and irregular word reading has been used to motivate the distinction between a sublexical reading route which stores regular grapheme-to-phoneme correspondences (GPCs) and hence can pronounce nonwords, and a whole-word lexical/semantic route which must be used for irregular words (e.g. Coltheart, 1985; Coltheart, Curtis, Atkins, & Haller, 1993; Morton & Patterson, 1980). Figure 1 shows one possible box and arrow account of the processes underlying reading and spelling, which embodies the putative distinction between lexical and sublexical routes.

In the light of the earlier discussion, the validity of the inference from double dissociation to a particular theory of the modular organisation of the cognitive system under study depends on (1) how well that theory predicts the double dissociation and (2) how well the other theories predict a double dissociation. The validity of (1) and (2), and hence how well double dissociations can distinguish between rival accounts of the functional organisation of the cognitive system, depends on the class of theories under consideration. Let us consider the situation where this class includes just box and arrow models, and where it also includes neural networks.

Boxes and Arrows

Any box and arrow model that has some component which is selectively used for task A and another which is selectively used for task B can predict a double dissociation given the standard assumption that brain damage corresponds to selective damage to a particular box or arrow. Thus point (1) is straightforward.

Point (2), however, is less clear-cut. First, there will be many different modular architectures which can lead to a double dissociation. All that is required is that for both tasks, there is some component involved only in that task. That there is such a component says nothing about the function of that component, and nothing about how it fits into the rest of the cognitive system. So, for example, it is prima facie consistent with the double dissociation between long- and short-term memory that the memory system consists of a very large and complex array of modules, all of which are shared between short- and long-term memory, except for two modules, one of

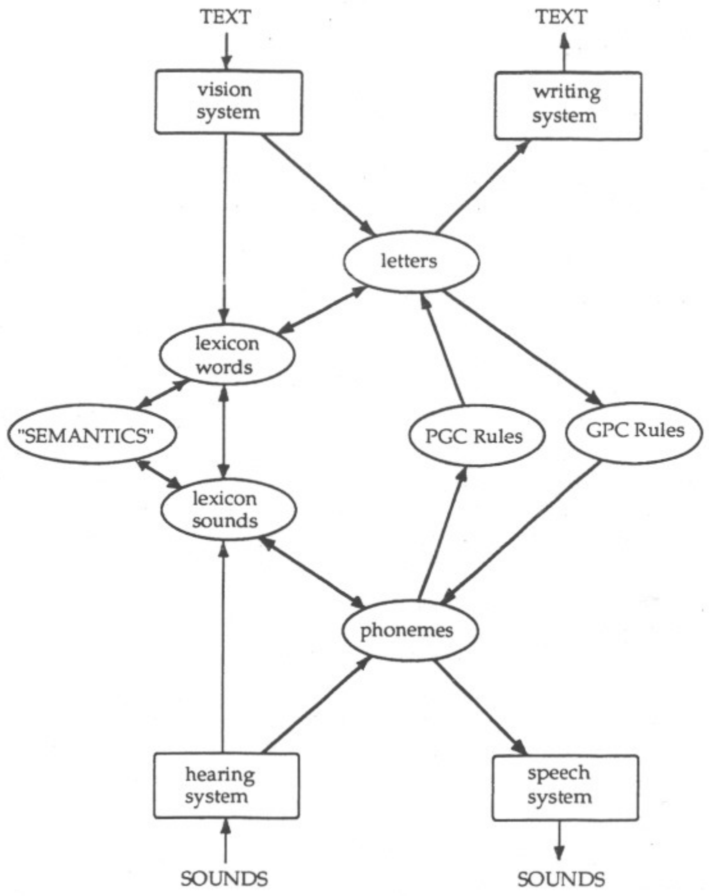

FIG. 1. One possible "box and arrow" account of reading and spelling based on the dual-route reading model of Coltheart et al. (1993).

which has some function or other specific to remembering information over long periods and one which has some function or other specific to remembering information over short periods. Second, double dissociations between two tasks can in principle occur even when there is no specific dedicated module for either task (Dunn & Kirsner, 1988; Shallice, 1988; see Chater & Ganis, 1991, for a simple illustrative example). Shallice (1988) notes, for example, that a processing continuum (such as is found in topographic maps in the visual cortex, in which regions of cortex correspond to regions of the visual field) can give rise to double dissociations, even though the visual cortex may not be divided up into isolable modules. Chater and Ganis (1991) show that damage to a simple electrical circuit can give rise to a double dissociation between two very simple binary switching tasks. While this system is made of simple, modular components, Chater and Ganis found that all of the components are involved in the normal performance of both tasks; that is, no module is task-specific, as would be expected according to the standard inference from double dissociation to modularity of function.

The claims concerning what can be learnt from double dissociations are often put more strongly, however. For example, Marin, Saffran and Schwartz (1976, pp. 869–870) state that: "At the very least ... [observed double dissociations] ... should yield a taxonomy of functional subsystems. It may not tell us how these subsystems interact—but it should identify and describe what distinct capacities are available". That is, they argue that double dissociations should specify the components of a box and arrow model of a cognitive system. As we have seen, such claims are not justified, even if consideration is limited to modular systems.

Neural Networks

Since the double dissociation inference is intended to map out, or at least constrain, the architecture of the cognitive system under study in terms of boxes and arrows, it might seem that neural network models are necessarily irrelevant to this aspect of neuropsychological methodology. Neural network models, the argument might go, are concerned with a level of detail below that of the box and arrow diagram. Most straightforwardly, they may be construed as models of the operation of particular components of such a model, a level of detail double dissociation does not aim to uncover. This line of reasoning suggests that cognitive neuropsychology can proceed without concern for neural network models of cognition (see, e.g. Kosslyn, Flynn, Amsterdam, & Wang, 1990, for this point of view). The reason that this line of argument is not convincing is that it does not consider the possibility that a single neural network, without any obvious box and arrow structure, might be able to produce double dissociations, which would mislead the cognitive neuropsychologist into postulating a modular structure where none was present.

So, to focus on the class of examples with which we will be concerned below, neural network approaches have frequently aimed to model rule-governed and rule exceptional behaviour in using a single network, whereas box and arrow models, such as that shown in Fig. 1, treat them as separate. Rumelhart and McClelland (1986), Pinker and Prince (1988), Pinker (1991), MacWhinney and Leinbach (1991) and Bullinaria (1994b) discuss the English past tense; Seidenberg and McClelland (1989), Besner, Twilley, McCann and Seergobin (1990), Coltheart et al. (1993), Plaut and McClelland (1993) and Bullinaria (1994a; 1994c) discuss reading; Brown, Loosemore and Watson (1994) and Bullinaria (1994b) discuss spelling. Hence, in terms of the discussion of neuropsychological methodology above, neural network models enlarge the class of theories concerning normal function. From the point of view of neuropsychological inference, the crucial question is what predictions do such models make about the patterns of breakdown. Focusing on double dissociations between performance on the processing of rule-governed and exceptional items, a central issue is whether a "single route" model of rule and non-rule behaviour can give rise to double dissociations. If it can, then the inference from double dissociation to modularity of function appears to be under threat, in this context at least; if it cannot, the traditional inference does not seem to be challenged by neural network accounts. We shall discuss this question and examine some relevant case studies below.

Wood (1978) and Sartori (1988) discuss simple demonstration simulations which seem to give rise to dissociation-like effects on simple pattern association tasks. Shallice (1988, p. 254), however, argues that these cases are not persuasive, since mere associations rather than independent tasks are dissociated and because the experiments are very small scale. Furthermore, he argues that the small scale of these experiments means that individual units and connections play an important role in the functioning of the system and this is unlikely to be true in more realistic neural networks, where function is distributed over a very large number of units and connections, so that no particular component of the system has a large influence on its overall behaviour. He goes on to conclude that "there is as yet no suggestion that a strong double dissociation can take place from two lesions within a properly distributed network".

In the light of these suggestive but inconclusive findings, an obvious next step is to conduct a systematic series of case studies on somewhat larger networks. Furthermore, given the centrality of inferences concerning performance on rule-governed and exceptional items for much neuropsychology, it is interesting to consider a simplified task domain in which there are two kinds of mapping, one which is in accordance with a rule and another which provides exceptions to that rule. We are therefore led to consider explicit case studies in which small neural networks are trained on such a pair of tasks, systematically lesioned and examined for evidence of dissociation between the tasks.

SMALL-SCALE NEURAL NETWORK SIMULATIONS

In this section, we begin by outlining some of the general problems encountered when attempting to simulate brain damage in artificial neural networks. We then describe two simple toy models that represent many features of the real data discussed previously and present some typical learning and damage results.

General Considerations

We noted earlier that the patterns of breakdown that may be predicted from a cognitive level account of function depend on, among other things, how that account is implemented in the brain, and hence how real brain damage is likely to affect the cognitive level. We also noted that, in standard box and arrow accounts, little is said concerning the relationship between cognitive and neural levels. The assumption, often implicit, is that boxes and arrows correspond to areas and pathways in the brain, and that damage to a localised brain region may thus result in the removal of one or more boxes and/or arrows at the cognitive level of description.

Neural network models of cognition promise, by contrast, to model different types of damage much more directly. By viewing artificial neural networks (ANNs) as (vastly oversimplified) counterparts of real networks of neurons, the patterns of damage between ANNs and real neural systems can be equated. Thus, removing a hidden unit from an ANN can be viewed as analogous to removing a neuron in a real neural system; removing a connection may be viewed as analogous to the removal of a synapse or dendrite; the rescaling of weights in an ANN can be viewed as analogous, for example, to change in the balance of neurotransmitters. But, of course, these analogies are very distant. The relevance of ANN simulations for neuropsychology depends on the assumption that the simplifications ANNs embody are not crucial with respect to the effects of damage: the effects must be supposed similar for any network-like system.

It is not possible to distinguish confidently aspects of behaviour which are common to all network-like systems, including the nervous system. However, it is possible to consider a range of different artificial cases, to see which aspects of behaviour under damage are consistent, and which vary capriciously, and to attempt to back up empirical findings with theoretical analysis. This at least allows more informed speculation about what might be expected from damage in the nervous system.

There is a wide range of factors that may be varied in artificial neural network simulations. The architecture of the network (e.g. feedforward *vs* recurrent networks, the number of hidden units, whether the network is fully or sparsely connected, and so on) may potentially affect the patterns of breakdown observed when the network is damaged. Also, it is possible that using different learning algorithms to train the network will result in different network configurations with different breakdown patterns. Furthermore, even when the architecture and learning algorithm are held constant, very different network configurations can result if a different representation of the problem is chosen (i.e. different representations of the input and output patterns). Highly localist input and output codes might, for example, be expected to induce different hidden unit representations in the network from very distributed input and output representations; and this in turn might have an impact on the selectivity of damage to those hidden units. Another factor that may potentially affect the results is the way in which the network is damaged—whether hidden units are removed, or connections removed, or noise is added to the weights on the connections, or the values of the weights are globally rescaled, and so on (e.g. Small, 1991). Finally, even if we chose a particular network architecture, trained and damaged in a particular way, on a fixed representation of the problem, it is still possible that a variety of different results might be obtained. This is because ANNs are notoriously dependent on the precise values of their parameters, such as the learning rate, and the particular small random weights assigned before training begins (e.g. Fahlman, 1988; Kolen & Pollack, 1991; Kramer & Sangiovanni-Vincentelli, 1989).

To be able to draw general conclusions with complete confidence, it would be necessary to explore extensively all of these sources of variation together. Since the number of combinations is enormous, we must limit ourselves, in practice, to exploring a very restricted subset of possibilities. For example, in the present study, we consider only the standard fully connected feedforward network architecture with one hidden layer, since this is by far the most widely used architecture in connectionist cognitive modelling. We shall discuss the implications of this restriction on pp. 260–261. This still leaves the important architectural question concerning the number of hidden units used. Specifically, should the network be near minimal—that is, have nearly the minimum number of units and connections necessary to solve the given problem? Or should a much larger number of hidden units than needed to solve the problem be used? Using near-minimal ANNs often speeds up the training, improves the generalisation and makes it easier to understand the representations that have been learnt in the hidden layer. In earlier work on such near-minimal systems (Bullinaria & Chater, 1993), we found that double dissociations can occur with the kind of rule/exception mapping tasks that we discuss in this paper. We shall see later, however, that with highly non-minimal network configurations such double dissociations are not found.

It would appear quite unlikely that results obtained from near-minimal network configurations are psychologically and biologically relevant. The number of units available to the brain is very large, and it is not clear how small numbers of units might be segregated out for the solution of particular problems. Furthermore, to obtain near-minimal configurations, the number of units required to solve a given problem must be known beforehand, and this information is not available when learning begins. Finally, near-minimal networks are not tolerant to damage, since even small lesions will result in a breakdown in performance. Thus, such networks do not match the brain's robustness to damage. For this reason, we concentrate on non-minimal networks in this paper. In particular, we shall start with networks with about ten times as many hidden units than are actually needed and then consider how our results change as we use more or less.

Closely related to our choice of architecture is the choice of representations. In ANN models of word-naming and past tense acquisition, for example, a wide range of different representations of the problem has been used, imposing very different demands on the learning system. Indeed, some such models would not work at all without their carefully chosen representations. For example, the original NETtalk model of reading (Sejnowski & Rosenberg, 1987) requires the training data to be pre-processed in order to align the letters and phonemes, and MacWhinney and Leinbach's (1991) past-tense learning model requires the training data to fit into a series of templates. In this paper, we bypass these problems of representation by choosing an abstract set of training data that mimics the important features found in many realistic problems.

Once we have decided on a particular network and training data, we have to choose a learning algorithm. Almost all connectionist cognitive models are trained by some variant of gradient descent learning, most notably back-propagation, and this kind of method is our focus here.[2] Within the class of gradient descent learning methods, previous studies (e.g. Plaut & Shallice, 1993) have indicated that the main results actually show very little learning algorithm dependence. In this study, we shall concentrate on a particular form of gradient descent learning, using the conjugate gradient algorithm (Kramer & Sangiovanni-Vincentelli, 1989). Finally, we need to check that our results do not depend significantly on the parameters used by the learning algorithm nor the different small random initial weights prior to

[2] The other main class of supervised connectionist learning procedures, which are seldom used in cognitive modelling, are constructive learning algorithms, in which the number of units involved in solving the problem is not fixed, but is itself determined by learning. For comparison, we initially conducted a number of experiments with perhaps the best known of these procedures, namely cascade correlation (Fahlman & Lebiere, 1990). These algorithms tend to produce near-minimal networks, which are not the focus of interest here. As we noted above, with fixed near-minimal networks trained by gradient descent, double dissociations between rules and exceptions are observed, and this applied also to networks generated using cascade correlation.

training. By using conjugate gradient learning with line searches, we avoid the well-known problems caused by using an inappropriate fixed step size for the gradient descent.[3] For different initial weights we do expect slightly different results, in the same way that different real brains show individual idiosyncrasies, but we will hopefully find consistent trends across all cases.

Finally, once trained, there are then many different forms of network damage that need to be considered (Small, 1991). The most obvious are the removal of subsets of units and connections, but we also need to consider the various possible changes to the weights and activations. We shall concentrate on five types of damage which should be fairly representative of the possibilities: (1) global reduction of weights by rescaling; (2) global reduction of weights by subtraction; (3) addition of Gaussian random noise to all weights; (4) removal of hidden units; and (5) removal of connections. Later we shall examine how relearning after damage can affect performance.

Learning and Damage

The forms of double dissociation that we shall investigate are between regularities and their exceptions, where the exceptions may consist of isolated examples, or may themselves be governed by sub-regularities. The degree to which the fundamental regularity has been grasped is assessed, in patients and in computational models, by testing whether the regularity can be generalised successfully to novel items. The degree to which the exceptions are grasped may be assessed directly, by testing performance on those exceptions. The question of interest is: Do people use separate mechanisms to deal with the regular and exceptional cases? As we discussed above, in the domains of reading and spelling, double dissociations have been taken to provide strong evidence for distinct mechanisms: surface dyslexics appear to show intact knowledge of the regularities, but impaired knowledge of the exceptions, whereas phonological dyslexics appear to show the opposite profile (e.g. Coltheart et al., 1993; Shallice, 1988). Closely related issues arise regarding rules and exceptions in morphological processing, such as the processing of verb inflections (e.g. Pinker, 1991).

Reading, spelling and verb inflection in English follow a whole series of rules, sub-rules and exceptions. In this section, we shall use artificial versions

[3] Conjugate gradient training operates by choosing an appropriate direction in weight space dependent upon the current gradient on the error surface and the previous step direction. A search along this direction is then carried out to find a step size that gives an acceptably near optimal decrease in error. There are necessarily free parameters that determine the accuracy to which this line search is carried out, but these are generally found to be much less crucial to the final network performance than a poor choice of fixed step size in an algorithm such as back-propagation.

of these "quasi-regular" mappings in which we have abstracted out their essential features: rules, sub-rules and exceptions. This allows us to examine the basic possibilities free of many of the problems discussed above and also make our results applicable much more widely. Later we shall see that our results do also hold for the more realistic mappings from which they were derived.

Thus, to begin with, we simulated populations of small feedforward networks, with one hidden layer, trained on two different quasi-regular mappings. The first mapping involved learning a rule and a less frequent sub-rule; the second involved learning a rule with a number of totally random exceptions. Each network was then lesioned in the five ways described above and the output performances determined for the two training data subsets and the corresponding generalisation test set. The performance in each case was measured as the percentage of output patterns produced correctly, so that the types and degrees of dissociations could be compared between different networks and tasks.

Each network was trained using the conjugate gradient learning algorithm (Kramer & Sangiovanni-Vincentelli, 1989) to reduce the sum-squared error on the output activations. Normally such networks will converge to a state where they have learned all the main regularities but few (if any) of the sub-regularities or exceptions. This is fine if the exceptions are merely noise in the training data which we generally do not want to learn anyway. If, as here, we do want to learn the exceptions as well, we need to use a different error measure (e.g. cross entropy as in Hinton, 1989) or modify the learning process (e.g. by introducing a sigmoid prime offset as in Fahlman, 1988). For the present study, we used a sigmoid prime offset of 0.1 during the early stages of training (for a discussion of the need to do this in more realistic models of reading and spelling, see Bullinaria, 1994a; 1994b). This, of course, slows down the learning because the line search direction is no longer optimal, but it eventually resulted in perfect performance on the training data in all our simulations. Experiments with different values of sigmoid prime offset and different "turning off" points indicate that the precise values do not qualitatively affect our main conclusions.

The first set of simulations involved networks with 9 input units, 100 hidden units and 9 output units that were trained to learn a rule and sub-rule. The training data consisted of the identity map, except that when the first four bits are 1111 or 0000, the last three bits are flipped (e.g. 101010101 → 101010101 but 111110101 → 111110010). A random subset of half the full set of 512 possible patterns was used for training (giving 224 "rules" and 32 "sub-rules") and the remaining 256 patterns were used for testing generalisation.

We trained 10 networks starting from different small random initial weights until the total output error score was less than 0.0001 (on p. 257 we discuss the effect of using different stopping criteria). Each network then had perfect performance on both the training and generalisation data. (An output unit was deemed to be producing the "correct" output if its activation error was less than 0.5, and an output pattern was deemed to be "correct" if each output unit was "correct".) The networks were then lesioned in the five ways described above with the level of damage gradually being increased in about 25 steps until the network failed to produce any correct output patterns at all. At each step, we calculated the difference between the percentage of rule items for which the output was incorrect and the percentage of sub-rule items for which the output was incorrect. This provides a measure of dissociation between performance on rule and sub-rule items.

Two cases must be considered, depending on the direction of the dissociation. First, the sub-rule items may be more impaired than the rule items. The strongest possible selective impairment of sub-rule items would score 100, meaning that 100% of sub-rule items are lost, and 0% of rule items are lost. Second, the rule items may be more impaired than the sub-rule items. The strongest possible selective impairment of rule items would also score 100, meaning that 100% of rule items are lost, and 0% of sub-rule items are lost.

For each network, the maximum dissociation in each of these two directions observed during each sequence of damage was recorded. These maximal dissociations were then averaged over 20 damage runs of each type for each of the 10 networks. Table 1 summarises the mean dissociations between rules and sub-rules. To give some idea of the range of values obtained, the strongest and weakest dissociations are also given. The figures in this table, and subsequent tables, are the differences in the percentage of items lost as described in the previous paragraph.

Table 2 shows the corresponding dissociations between the generalisation and sub-rule performance (i.e. the difference in performance on the generalisation test set compared with that on the sub-regular training items). In psychological terms, rule performance corresponds to performance on the most regular items; sub-rule performance corresponds to performance on irregulars (specifically, those irregulars which are covered by a sub-regularity, rather than being completely irregular); generalisation performance corresponds to grasp of the rule(s) governing the stimuli.

The results show clearly that dissociations can occur in both directions. This shows that double dissociations can occur even in networks with many more hidden units than are actually needed. But the dissociations are stronger where the sub-rule is selectively lost, than where the rule is selectively lost. This difference is particularly clear in the cases of global weight changes.

TABLE 1
Rule/Sub-rule, 100 Hidden Units

Form of damage	Sub-rules lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	90.6	88.7	84.4	0.0	0.7	3.6
Global reduction of weights	97.8	86.5	75.0	0.0	0.0	0.0
Adding noise to weights	74.6	42.0	7.1	0.0	2.9	17.4
Removing hidden units	74.6	31.9	2.2	0.0	13.1	51.3
Removing connections	87.5	45.1	2.2	0.0	12.4	38.9

TABLE 2
Generalisation/Sub-rule, 100 Hidden Units

Form of damage	Sub-rules lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	90.6	87.8	82.6	0.0	0.6	3.1
Global reduction of weights	95.5	85.3	73.7	0.0	0.0	0.0
Adding noise to weights	70.1	39.3	7.6	0.0	3.7	22.8
Removing hidden units	73.2	29.8	0.9	0.0	14.0	46.9
Removing connections	86.6	43.8	0.4	0.0	12.9	37.9

TABLE 3
Rule/Exception, 100 Hidden Units

Form of damage	Exceptions lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.0
Global reduction of weights	89.6	77.9	71.2	0.0	0.0	0.0
Adding noise to weights	91.2	67.5	42.1	0.0	0.1	3.8
Removing hidden units	63.8	39.1	7.9	0.0	3.7	25.8
Removing connections	88.8	57.9	20.8	0.0	0.1	8.8

TABLE 4
Generalisation/Exception, 100 Hidden Units

Form of damage	Exceptions lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	99.7	99.1	0.0	0.1	0.5
Global reduction of weights	82.9	71.3	55.9	0.4	3.7	7.3
Adding noise to weights	86.7	61.4	34.1	0.0	3.0	15.7
Removing hidden units	57.6	31.3	21.7	0.0	7.2	25.4
Removing connections	79.2	52.1	22.1	0.0	1.1	12.9

The second set of simulations also involved networks with 9 input units, 100 hidden units and 9 output units, but here they were trained to learn a rule with random exceptions. The training data consisted of the identity map, except for 16 random input patterns which each mapped to a random output pattern. A random subset of half the full set of 512 possible input patterns was used for training (giving 224 "rules" and 16 "exceptions") and the remaining 256 input patterns were used for testing generalisation.[4] The lesioning procedure was the same as above and the resulting dissociations are shown in Tables 3 and 4. Again we find double dissociations and again the magnitude of the dissociations in which the rule is lost is smaller than the reversed dissociations. In fact, this asymmetry is significantly larger here than in the previous simulations. Also, we see that the generalisation results again mirror closely the training data results. Indeed, this mirroring applies not just on average, but can also be seen in individual damage runs.

We noted above that early small-scale neural network simulations (e.g. Sartori, 1988; Wood, 1978) were criticised as unrealistic because they are not fully distributed (Shallice, 1988). A particular problem that has been observed in the past with simulations such as ours is that the results can be obscured by dissociations that are merely artefacts of the input and output representations (e.g. Bullinaria & Chater, 1993; Shallice, 1988). If a single input or output unit on its own has an effect on the network output, then we are at the very least in violation of the assumption of full distribution. If the state of that single unit also has a significant determining effect on the class of the whole pattern, then damage to that unit can also result in artefactual dissociations.

We have chosen our toy problems carefully to minimise these problems, but to check this point we repeated the above simulations using more distributed input and output codings. We did this by replacing each "0" bit in the training data by the four bits "0011" and each "1" by "1100". Apart from the obvious increase in the number of input and output units from 9 to 36, this also makes the codes error-correcting (e.g. Dietterich & Bakiri, 1991). If the output of the network is deemed to be whichever of "0011" or "1100" is closest to the actual network output activations, then any one of the network output bits can be completely wrong without affecting the overall output.

Tables 5 and 6 show the dissociation results for the rule/sub-rule task with our new distributed representation and Tables 7 and 8 show the corresponding results for the rule/exeption case. From these we see that, for all types of damage apart from hidden unit removal, the dissociations with

[4] The 16 exceptional training patterns were: 000000111, 001001111; 111111101, 000101011; 100100011, 110000011; 100001100, 101000101; 011001100, 001000011; 111001101, 001010100; 100100101, 000111011; 111010001, 011010101; 101111100, 111010010; 110100110, 011101111; 000011001, 110110100; 110111010, 101011100; 100100110, 111111001; 000011010, 100001001; 011110000, 110011110; 101101001, 000001110.

TABLE 5
Rule/Sub-rule, Distributed Representation, 100 Hidden Units

Form of damage	Sub-rules lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.0
Global reduction of weights	100.0	95.6	83.0	0.0	0.0	0.0
Adding noise to weights	86.2	56.7	20.5	0.0	0.8	16.1
Removing hidden units	76.3	33.9	0.9	0.0	14.4	48.0
Removing connections	100.0	71.2	29.9	0.0	1.0	19.2

TABLE 6
Generalisation/Sub-rule, Distributed Representation, 100 Hidden Units

Form of damage	Sub-rules lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.1	0.9
Global reduction of weights	100.0	95.5	79.0	0.0	0.2	1.8
Adding noise to weights	82.1	53.9	22.8	0.0	1.8	17.4
Removing hidden units	77.7	30.1	7.5	0.0	16.3	55.8
Removing connections	100.0	68.6	29.9	0.0	2.3	20.1

TABLE 7
Rule/Exception, Distributed Representation, 100 Hidden Units

Form of damage	Exceptions lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.0
Global reduction of weights	100.0	96.6	91.7	0.0	0.0	0.0
Adding noise to weights	98.3	87.6	66.2	0.0	0.0	1.2
Removing hidden units	62.5	35.3	3.8	0.0	4.9	28.3
Removing connections	99.6	89.0	71.7	0.0	0.0	0.4

TABLE 8
Generalisation/Exception, Distributed Representation, 100 Hidden Units

Form of damage	Exceptions lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.4
Global reduction of weights	96.6	91.3	83.5	0.8	2.1	3.3
Adding noise to weights	94.2	80.2	58.9	0.0	2.9	9.1
Removing hidden units	50.2	24.7	0.4	0.0	11.6	34.5
Removing connections	94.3	82.0	57.7	0.0	2.9	9.4

the sub-rules/exceptions lost have been increased while those with the rules/generalisation lost have been decreased (we shall examine exactly why this happens on pp. 252–257). The lack of changes for the hidden unit removal case is easily understood. The network weights are invariably highly correlated within each block of four input or output units. Thus the effect of the removal of each hidden unit has essentially the same effect as in the less distributed case.

These preliminary simulations appear to suggest that double dissociations between rules and exceptions can occur in a distributed system which does not have separate routes for dealing with rule-governed versus exceptional items. This appears to indicate that the inference from double dissociation to modularity of function may not be reliable for this kind of case. But, as we shall see, this impression is misleading, since this pattern of results is not repeated when larger, more realistic networks are used.

SCALING UP AND DOWN

We have found that dissociations are surprisingly common in our small-scale neural networks and that double dissociations can be found within a single network. This appears to support the claims of those who have attempted to cast doubt on the inference from double dissociation to modularity of function (e.g. Chater & Ganis, 1991; Dunn & Kirsner, 1988; Ganis & Chater, 1991; Sartori, 1988; Wood, 1978). However, as we shall see, a more detailed investigation suggests otherwise. In this section, we present the results of further simulations which indicate that the double dissociations disappear as we scale up to more realistically sized networks. In the next section, we attempt to understand the underlying causes of the dissociations we find, and then we investigate the effects of relearning after damage.

The Toy Models

The first important feature to note about our toy models is that the performance of a typical small-scale network does not generally degrade smoothly as we increase the degree of a particular type of damage. For example, Fig. 2 shows how the performance typically deteriorates as we remove random hidden units from a distributed rule/sub-rule network of the type described above. We not only see that the curves are far from smooth, but also that many of the largest dissociations are due to the random fluctuations in these curves. In fact, it is easy to check that, despite the network having many times the number of hidden units required to learn the data, there are still some units that on their own have a significant effect on the outputs. Consequently, the neural network cannot be considered to be "fully distributed" in the sense of Shallice (1988). This finding has important implications for all forms of network damage: As we increase the number of hidden units, it is possible that the network will become more distributed, causing all the damage curves to smooth out and leave us with only a single dissociation or with no dissociation at all.

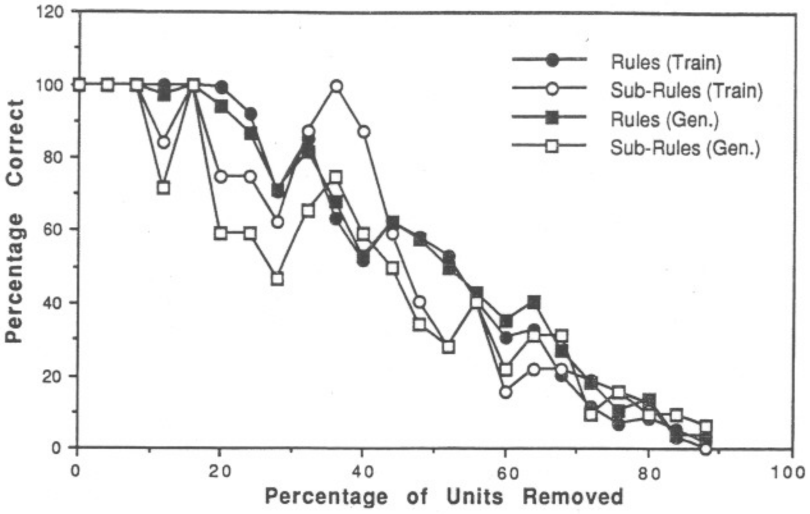

FIG. 2. Typical performance fall off due to damage by hidden unit removal of the distributed rule/sub-rule network with 100 hidden units.

Interestingly, we have found that damage by weight-scaling always gives a dissociation with the sub-rule or exceptions lost preferentially. While weight-scaling has a potentially direct neural interpretation, as mentioned above, in terms of changes in levels of neurotransmitters, this finding is perhaps of more interest because it can be viewed as indicative of what will be found as we scale up to very large, very distributed networks. The idea is that in a highly distributed network with a very large number of hidden units, removing large numbers of random connections or units will effectively have a global scaling effect on other parts of the system (since the specific effects of particular aspects of the damage will tend to average out). We may therefore expect that this form of single dissociation will be all we get as we scale up to larger networks. As we shall see, this expectation appears to be well founded.

Each output unit activation is given by the sigmoid of the sum of the contributions (i.e. activations × connection weights) from each hidden unit. The output is assumed to be "on" if its value is above 0.5, and "off" otherwise. The former occurs if the sum of contributions (including the threshold) is positive, and the latter occurs if the sum of contributions is negative. This means that the effect of damage on a network is only significant when the sum of contributions to one or more output units changes sign. At this point, a previously correct output will be judged to be incorrect.

This observation suggests a measure of the "vulnerability" of an output unit's performance on a given training pattern—the absolute magnitude of individual contributions to the unit, as a fraction of the absolute magnitude of the sum of contributions to that unit (Sanger, 1989). To see why this measure is useful, consider the following two cases. At one extreme, the output may be composed of many small, and roughly equal, contributions, each having the same sign. In this case, removing any particular hidden unit cannot significantly change the output—if there are many small contributions which are, say, positive, then the output will still be positive even if one or more of these is removed. At the other extreme, the output may be the average of many contributions, of large size, which cancel out because some are positive and some are negative. In this case, the absolute value of the individual contributions might even be larger than the absolute value of the sum of the contributions (i.e. the proportion of the total would be greater than 1). If hidden units associated with these large contributions are removed, then the sum of contributions may change sign, so that the output unit will change its response.

Naïvely, we would expect each individual contribution to become less important as we scale up to more hidden units. The larger the number of individual contributions we sum together, the less important we might expect any given individual contribution to be. Specifically, we would predict that the individual contributions will become smaller as a proportion of the total.

To test this requires a specific measure of the importance of individual contributions for a network relative to a given set of training patterns. A natural measure is the percentage of training patterns for which some output unit receives an individual contribution greater than a specific proportion of the total. If a network is highly distributed, then few training patterns will involve contributions which are a large proportion of the total contribution to any output unit. Therefore, performance on these training patterns is likely to be reasonably robust to damage. If a network is poorly distributed, then large contributions may be important for many training patterns, and hence such patterns will be vulnerable to damage. Specifically, the loss of an individual contribution over 1.0 of the total can change the output unit from "on" to "off" or vice versa; hence performance on that pattern will no longer be correct. Similarly, the removal of two contributions of 0.5, and more than three contributions of 0.3, could have the same effect. The number of patterns which contain these large contributions are therefore a measure of the "vulnerability" of these patterns to damage in which one, two or more hidden units may be removed.

Figure 3 shows a typical plot of the percentage of patterns dependent on individual contributions greater than 1.0, 0.5 and 0.3 of the total contribution for some output unit, plotted against the number of hidden units in the network. Rule and sub-rule patterns are plotted separately. We see that, for our rule/sub-rule problem with distributed representations, the number of important contributions do indeed decrease in the expected manner with the number of hidden units. In particular, we see that we need more than 150 hidden units to ensure that no single hidden unit has a significant effect on the output patterns, and to be sure that no unlucky combinations of two or three units have a significant effect, we need to go beyond 600 hidden units. We also see that the sub-rule patterns are more prone to having large contributions than the more regular items. A similar set of curves are found for the rule/exception case, except that there the size of contributions falls off faster with the number of hidden units. Clearly, the details of such curves will be rather problem (and learning algorithm) dependent and would have to be checked explicitly for each case.

We have argued above that brains presumably allocate many more hidden units (neurons) to solving a problem than are actually required, so we may expect them to be operating in a region where all the individual contributions are very small and all the damage curves are very smooth. Thus, for our simulations to be an acceptable approximation to what is happening in real brains, we need to check that our networks are operating in a similar regime. Our results, such as those shown in Fig. 3, suggest that this can mean having to use tens, if not hundreds, of times more hidden units than are actually required to solve the problem.

Tables 9 and 10 show the dissociations obtained for the distributed rule/exception networks when we increase the number of hidden units to 600. We see that the dissociations with the rules lost have now virtually disappeared (compared with the corresponding 100 hidden unit case of Tables 7 and 8). The only instances remaining are caused by random fluctuations in the hidden unit removal damage curves at relatively late stages of damage (when the number of hidden units has become quite small again and hence individual contributions are more crucial). For example, Fig. 4 shows how a typical large dissociation (of 13.3%) with predominantly rules lost arises after 80% damage (corresponding to 480 out of 600 hidden units removed) in a run otherwise dominated by a dissociation with the exceptions lost. As before, the generalisation/exception results mirror the rule/exception results. As we increase the number of hidden units further, we will be left with only the single dissociation with the least regular patterns lost first.

Finally, to confirm our intuitions, we consider decreasing the number of hidden units to 40. Tables 11 and 12 show that we do indeed find the

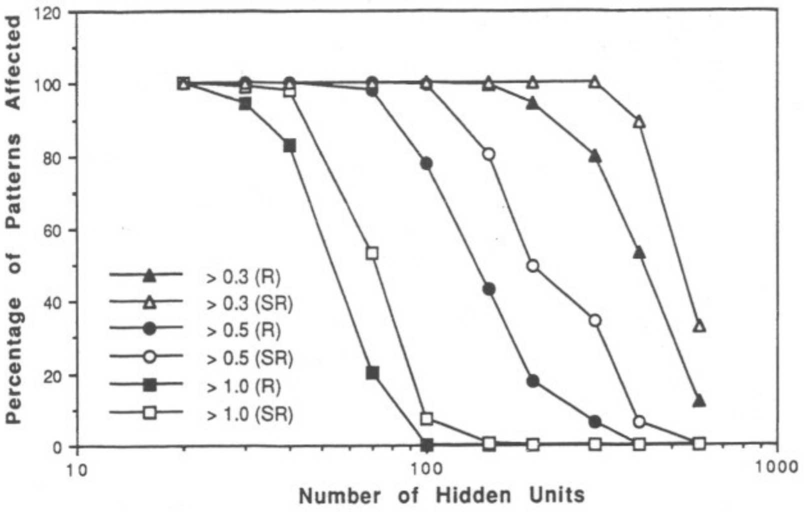

FIG. 3. The number of training patterns dependent on contributions greater than 0.3, 0.5 and 1.0 of the total versus the number of hidden units in the network, for the rule/sub-rule problem with distributed representations.

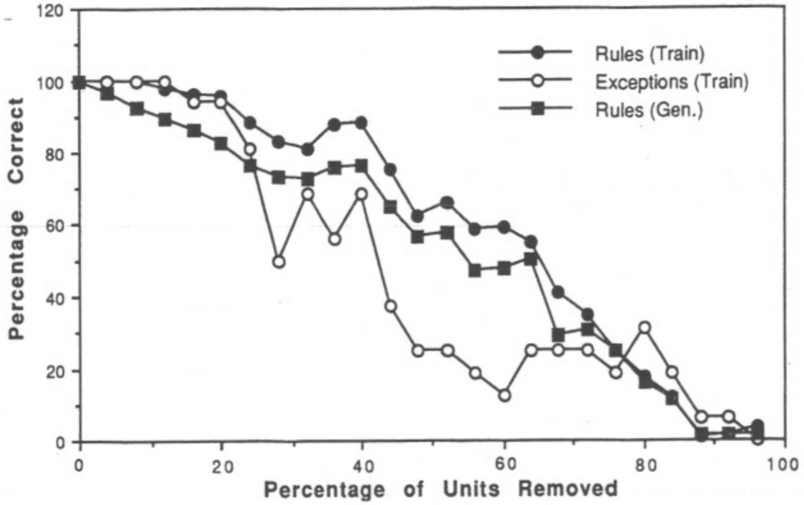

FIG. 4. Typical performance fall off due to damage by hidden unit removal of the distributed rule/exception network with 600 hidden units.

TABLE 9
Rule/Exception, Distributed Representation, 600 Hidden Units

Form of damage	Exceptions lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.0
Global reduction of weights	69.6	65.8	61.7	0.0	0.0	0.0
Adding noise to weights	94.2	83.2	57.5	0.0	0.0	0.4
Removing hidden units	69.2	39.6	17.9	0.0	3.3	17.5
Removing connections	100.0	96.0	86.2	0.0	0.0	0.0

TABLE 10
Generalisation/Exception, Distributed Representation, 600 Hidden Units

Form of damage	Exceptions lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	100.0	100.0	0.0	0.0	0.0
Global reduction of weights	61.4	58.6	53.1	0.8	2.1	3.9
Adding noise to weights	91.8	76.7	51.0	0.0	0.9	6.0
Removing hidden units	57.5	31.0	14.5	1.7	11.2	24.5
Removing connections	100.0	91.5	77.4	0.4	2.1	4.5

TABLE 11
Rule/Sub-rule, Distributed Representation, 40 Hidden Units

Form of damage	Sub-rules lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	93.2	80.4	0.0	0.0	0.0
Global reduction of weights	100.0	86.5	67.9	0.0	0.3	2.7
Adding noise to weights	80.8	51.1	14.3	0.0	2.7	25.4
Removing hidden units	59.4	27.8	0.9	0.0	11.4	43.8
Removing connections	90.6	58.0	23.7	0.0	3.7	38.8

TABLE 12
Generalisation/Sub-rule, Distributed Representation, 40 Hidden Units

Form of damage	Sub-rules lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	100.0	89.1	76.3	0.0	4.9	10.3
Global reduction of weights	100.0	84.9	61.6	0.0	4.9	15.2
Adding noise to weights	77.2	44.8	3.6	0.0	8.6	19.2
Removing hidden units	51.3	21.6	0.0	0.0	14.5	47.3
Removing connections	91.1	54.0	10.3	0.0	6.7	51.8

TABLE 13
Rule/Exception, Reading Model, 300 Hidden Units

Form of damage	Exceptions lost			Rules lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	54.2	52.1	50.0	2.1	2.1	2.1
Global reduction of weights	47.9	46.8	45.8	0.0	0.0	0.0
Adding noise to weights	50.0	34.5	18.8	0.0	0.3	2.1
Removing hidden units	52.1	27.1	12.5	0.0	4.9	18.8
Removing connections	56.2	35.6	18.8	0.0	3.3	12.6

TABLE 14
Generalisation/Exception, Reading Model, 300 Hidden Units

Form of damage	Exceptions lost			Generalization lost		
	Max	Mean	Min	Min	Mean	Max
Global scaling of weights	49.2	47.0	44.8	2.3	2.4	2.6
Global reduction of weights	41.4	38.6	35.7	2.3	3.5	4.7
Adding noise to weights	57.3	33.1	18.9	0.2	4.8	16.4
Removing hidden units	45.8	26.9	8.6	1.2	5.5	12.5
Removing connections	50.3	32.0	8.1	0.0	7.5	25.4

expected increase in strength of the double dissociations (compared with the corresponding 100 hidden unit case of Tables 5 and 6).

The pattern of results obtained with properly distributed, and reasonably large-scale, networks constrasts with that found with small, less distributed networks. As the size of the network increases, and the network becomes increasingly distributed, the pattern of double dissociation gives way to single dissociation: exceptions are lost while rules are preserved, but not the other way round. If the brain does indeed use highly distributed representations over very large numbers of units, then we should not expect double dissociations between rules and exceptions from a single system. So the inference from double dissociation to modularity of function may, with regard to rules and exceptions, be more reliable than small-scale network simulations would suggest.

More Realistic Models

We have argued that our toy models capture the essential features of a class of more realistic processes. In this sub-section, we summarise the results obtained from some more realistic models that confirm this claim.

The most well-known realistic example of a quasi-regular mapping is that between the orthographic and phonological representations in a language such as English. The double dissociation of interest there is that observed between performance on novel and irregular items in reading and spelling (e.g. Coltheart et al., 1993; Shallice, 1988). In reading, the double dissociation is evidenced by the dissociations found in patients with surface dyslexia (i.e. selective impairment of reading irregular words) and patients with phonological dyslexia (i.e. selective impairment of the rules as evidenced by poor reading of nonwords). In spelling, the double dissociation is evidenced by the corresponding dysgraphic syndromes.

Bullinaria (1994a) discusses a simple model of reading aloud (i.e. text-to-phoneme conversion). It is essentially a NETtalk style model (Sejnowski & Rosenberg, 1987) with a modified learning algorithm that obviates the need to pre-process the training data. It uses a simple localist representation for its inputs (i.e. one unit for each of 26 letters) and outputs (i.e. one unit for each of 38 phonemes including a phonemic null). The complete input layer consists of a moving window of 13 blocks of 26 units and the output layer consists of 2 blocks of 38 units. The most highly activated unit of each output block gives the phonemes corresponding to the letter activated in the central input block in the context of the letters in the other input blocks. Numerous variations of this basic system are discussed in detail in Bullinaria (1994a). We shall concentrate on two such networks with a single hidden layer of 300 units trained by back-propagation on the extended Seidenberg and McClelland (1989) corpus of 2998 monosyllabic words. These networks both achieved 100% performance on the training data and averaged 98% generalisation performance on a standard set of nonwords.

To test the effects of damage, we used the exception words and regular controls of Taraban and McClelland (1987) and the regular nonwords of Glushko (1979). Tables 13 and 14 summarise the dissociations obtained by following the same lesioning procedures as used for our toy models.

We see that the pattern of dissociation is very similar to those obtained above, with the surface dyslexic (exceptions lost) dissociations most pronounced. Some fairly large dissociations with the rules/generalisation lost do occur, although these do not approach the large dissociations found in patients with acquired phonological dyslexia. A closer examination of the damage curves indicates that these instances invariably correspond to major fluctuations in the damage curves, often when the overall performance is significantly degraded. On the basis of our abstract studies, we would predict that as the number of hidden units in such a model is increased further, single dissociations with rules preserved and exceptions lost would come increasingly to predominate.

A further indication of the generality of these findings is that similar patterns of performance after damage are observed in related models of spelling (Bullinaria, 1994b) and past tense acquisition (Bullinaria, 1994d).

UNDERSTANDING THE DISSOCIATIONS

In this section, we attempt to understand the underlying causes of the dissociations we find in our networks. By studying these causes, we can hope to gain a better conception of their generality and hence their potential relevance to the interpretation of neuropsychological data.

An analysis of the internal representations (i.e. patterns of hidden unit activations) learnt by the reading model (Bullinaria, 1994c) gave a fairly clear understanding of how various forms of damage to that model all result in symptoms similar to acquired surface dyslexia (with exceptions lost more than the regulars), whereas nothing results in anything like acquired phonological dyslexia (with the generalisation lost but not the training data). In this section, we shall abstract out the essential features of this analysis and apply it to our toy models.

The network's output weights project out particular directions in hidden unit activation space corresponding to each output unit. For each input pattern, each such projection is simply the sum of the contributions discussed above (i.e. the network's output before being passed through the sigmoid). During the training process, the network simultaneously learns these projection vectors and assigns each input pattern a point in hidden unit activation space with large positive or negative projections corresponding to output activations of "1" or "0". By plotting these projections for a typical output unit during learning and damage, we can gain a useful insight into what is happening.

During the early stages of training, the main regularities dominate the weight changes resulting in the exceptions being over-regularised. As the training proceeds, the exceptions are eventually learnt correctly, but they remain with higher error scores (i.e. less binary activations and smaller projections) than the regular items. Figure 5 shows a typical set of learning curves, the sum of the contributions to one particular output unit for 16 representative training patterns.

Despite the similarity between the dissociation patterns we observe, the corresponding damage curves are somewhat different for the different types of damage. We shall look at various typical damage plots for the 600 hidden unit distributed rule/exception networks. The corresponding plots for networks with less hidden units are similar but with larger statistical fluctuations.

Figure 6 shows the effect of global weight-scaling which results in very smooth curves that are essentially like the training curves in reverse. It is thus easy to understand how the strong (exceptions lost) dissociations occur here and why we never see the reversed dissociation. The damage trajectories due to global weight reductions by constant amounts are merely a rough approximation of this.

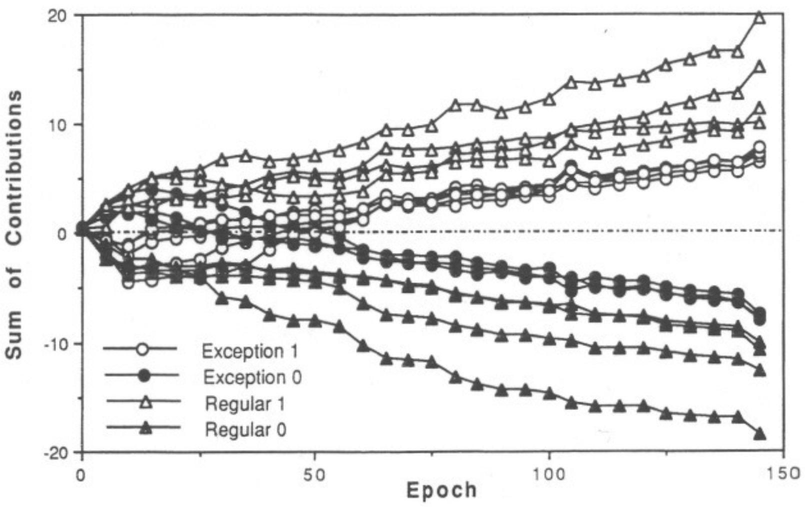

FIG. 5. Typical learning curves for a rule/exception network.

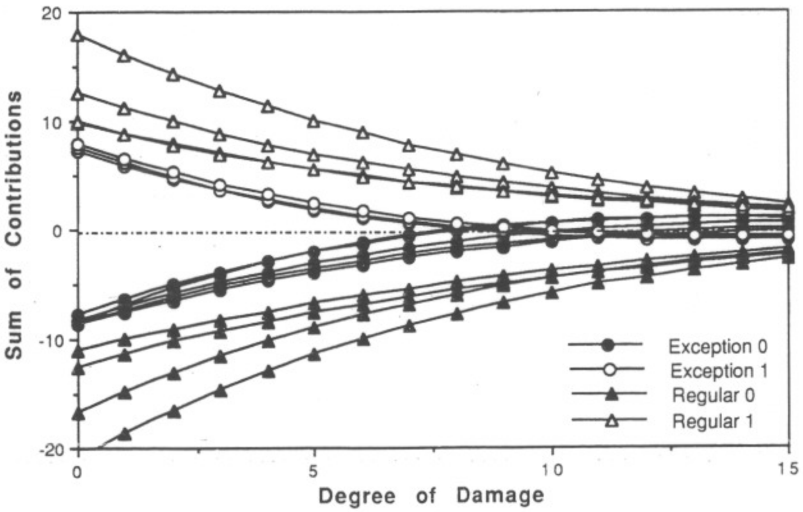

FIG. 6. Typical damage by weight-scaling plots for a rule/exception network.

In a network with many hidden units and where the contributions to each unit's activation are all very small and insignificant, removing a random subset of connections will be a good approximation to global weight-scaling, so we can expect similar damage plots.[5] This is confirmed explicitly by Fig. 7. We can thus also understand how the dissociations arise with this kind of damage.

It is clear from our results thus far, that the same simple argument will not apply to the removal of hidden units. In this paper, we have considered only ANNs with a single hidden layer. Can we draw any conclusions from these results for networks in which this biologically highly artificial restriction has been removed? In a multi-layer ANN, it seems reasonable to expect that, as we have found with removing connections in a single hidden layer network, removing random sets of hidden units would also approximate global weight-scaling and give similar damage plots. However, in a single hidden layer network, hidden unit removal is equivalent to scaling only the output weights, which is equivalent to merely squashing the outputs (since all the thresholds are small). Thus all the errors are due to fluctuations away from the smooth case, as seen in Fig. 8. Since the exceptions/sub-rules start off nearest the cross-over line, it is not surprising that they exhibit most errors and hence we get our single dissociation. Since the errors here necessarily arise from the fluctuations, it is easy to see why the dissociations do not scale as clearly as with the other forms of damage.

Damage by adding noise to the network weights gives a fundamentally different set of plots,[6] as shown in Fig. 9. Again the errors are predominantly the exceptions/sub-rules simply because the curves are fairly laminar and these cases start off nearest the cross-over line.

We have seen, for each type of damage, the importance of the regularity of the training patterns. Our explanation of what is happening implies that the greater the regularity difference between the two tasks, the greater the accuracy of the single dissociation. This is confirmed by the trend from double to single dissociation as we went from the sub-rule training data of Tables 1 and 2 to the true exceptions training data of Tables 3 and 4.

[5] Suppose the total input into an arbitrary unit is composed of N contributions $\{x_i\}$ distributed with mean X and maximum $x_{\max}$ much less than the total contribution NX. If N is sufficiently large, then (to good approximation) randomly removed contributions will follow the same distribution as the full set of contributions and hence leave the mean contribution X constant while reducing N to N' (at least if N' is still large). The total contribution is thus scaled from NX to $N'X = (N'/N)NX$, which is equivalent to simply scaling the weight in each contribution by (N'/N).

[6] In the notation of Footnote 5, to first approximation, both the mean contribution X and number of contributions N remain constant. All the changes are due to random fluctuations *not* averaging out.

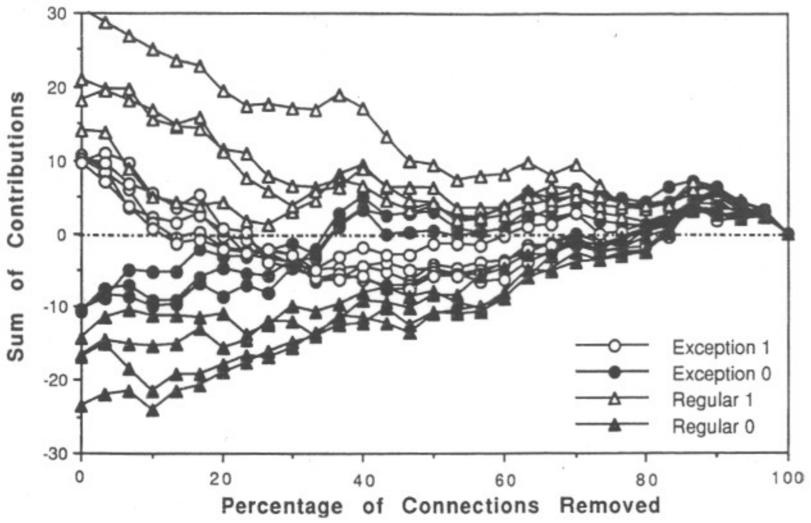

FIG. 7. Typical damage by connection removal plots for a rule/exception network.

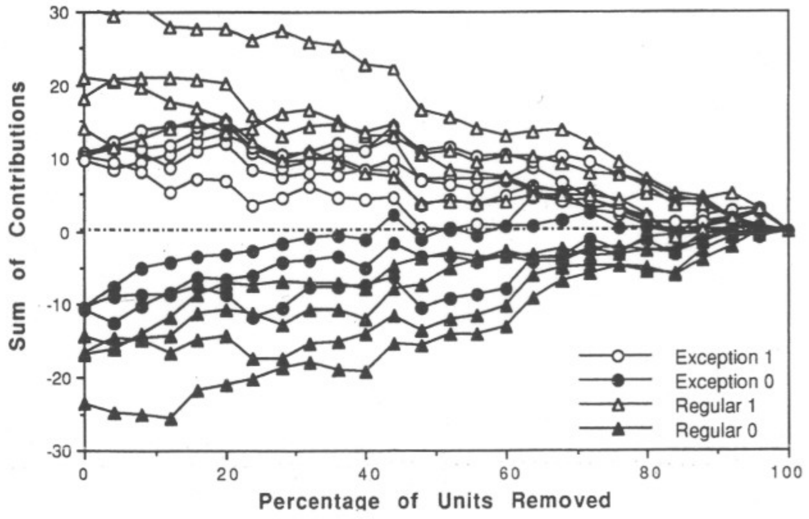

FIG. 8. Typical damage by hidden unit removal plots for a rule/exception network.

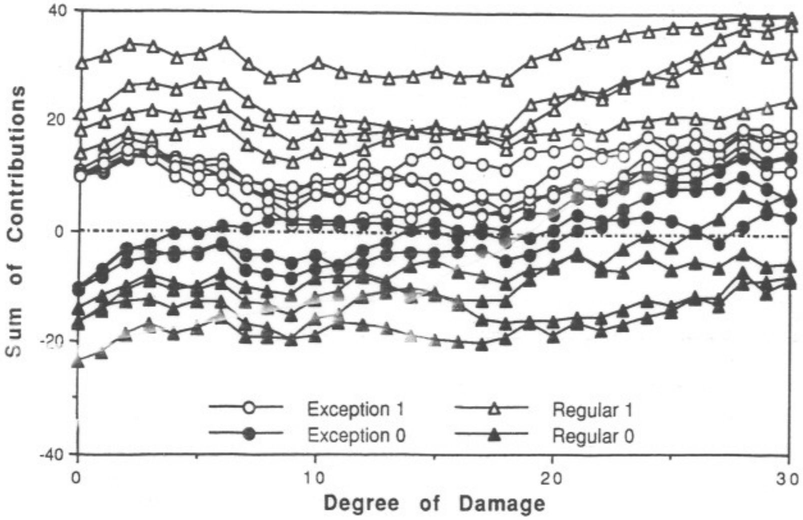

FIG. 9. Typical damage by noise addition plots for a rule/exception network.

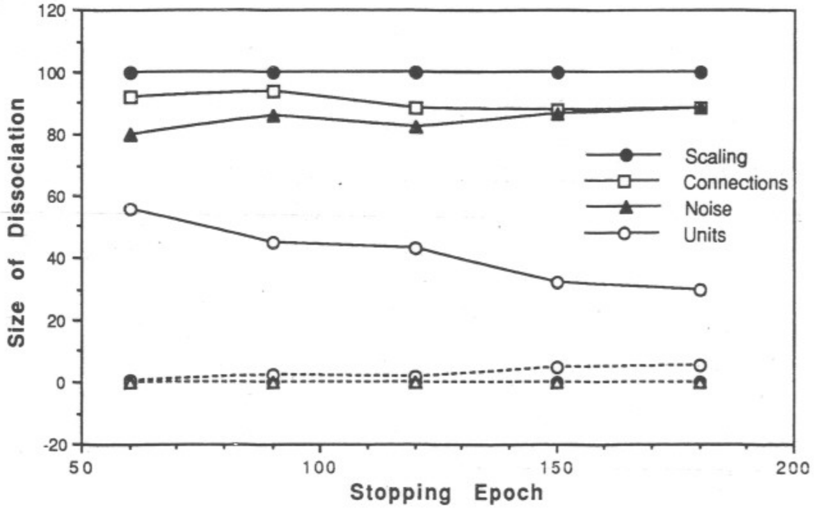

FIG. 10. Dependence of mean dissociations after damage on the amount of training originally received by a rule/exception network.

Our conclusion here seems to be that we can understand how damage to large-scale neural networks will give rise to a single dissociation determined by the regularity in the training data. This is simply because the network learning algorithm naturally acquires the regularities in their training data earlier and more strongly. It seems unlikely that large homogeneous feedforward networks will ever produce dissociations that are not simply a consequence of the regularity of their training data.

Now that we understand what is happening during training and damage, we are in a position to assess the effect of changing the stopping criterion for the network training. We can see from Fig. 5 that once all the exceptions have been learnt (around epoch 60), the distribution of total contributions remains qualitatively the same during further training. The effects of damage will also be qualitatively the same. The only difference will be the relative significance of random fluctuations in the damage. Consequently, we should expect the dissociations we find to be largely independent of the stopping criteria. This is easily checked explicitly. Figure 10 shows the mean dissociations obtained for the network of Fig. 5 with the training stopped at various stages (solid lines for the exceptions lost dissociations, dashed lines for the rules lost dissociations). We see that our main conclusions are indeed independent of the stopping criterion. The only major trend as we increase the amount of training is the slight reduction in size of the exception lost dissociation obtained by hidden unit removal.

In the reading model, the output representation is more complicated, with the network outputs determined by competition within each block of output units. Nevertheless, we come to similar conclusions there (for details of the corresponding analysis in that case, see Bullinaria, 1994c).

RELEARNING AFTER DAMAGE

In this section, we consider the issue of relearning after damage, which is often an important factor in determining the performance of neuropsychological patients. It is possible that, even if rule/exception double dissociations do not occur when full distributed networks are damaged, such dissociations could emerge when the network is retrained. We present simulations which suggest that relearning does not provide a mechanism for generating such double dissociations.

Neurological patients often (but not always) show a rapid improvement in performance after a lesion occurs (Geschwind, 1985; Plaut, in press a). This phenomenon has a ready explanation in connectionist terms. Given that there will still be a large amount of information left in the network even after damage, it should not be surprising that relearning after damage proceeds at a much faster rate than the original learning process (Hinton & Sejnowski, 1986; Sejnowski & Rosenberg, 1987). This can be seen clearly in Fig. 11, which shows the relearning curves after 300 hidden units are removed from one of our 600 hidden unit distributed rule/exception networks. The damage leaves the network with 75.8% performance on the regular training data and 37.5% on the exceptions (i.e. a 38.3% dissociation). After only six epochs of relearning, the network regains perfect performance on all the training data. This would typically take about 150 epochs during the original learning stage.

Aside from its intrinsic interest, the phenomenon of rapid relearning in patients and in networks might potentially upset the conclusions that we have drawn so far. Double dissociations are typically observed in patient populations where there has been significant time during which relearning may have occurred. Hence, the performance of such patients may perhaps be better modelled not by networks which are merely trained and damaged, such as those we have considered so far, but by networks which are retrained after damage. Hence, to establish our conclusion that double dissociations between rules and exceptions do not occur in fully distributed networks, we need to be confident that this also applies to networks that have experienced some retraining.

To see why this should be so, let us reconsider the initial pattern of learning, as shown in Fig. 5. The regularities dominate the training data, so when there are significant errors on them, they dominate the weight changes and drag the exceptional patterns with them. It is only when the regularities are learnt reasonably well (i.e. when they contribute less to the total network error score than the relatively few exceptions) that the weight changes can begin to work on getting the exceptions right.

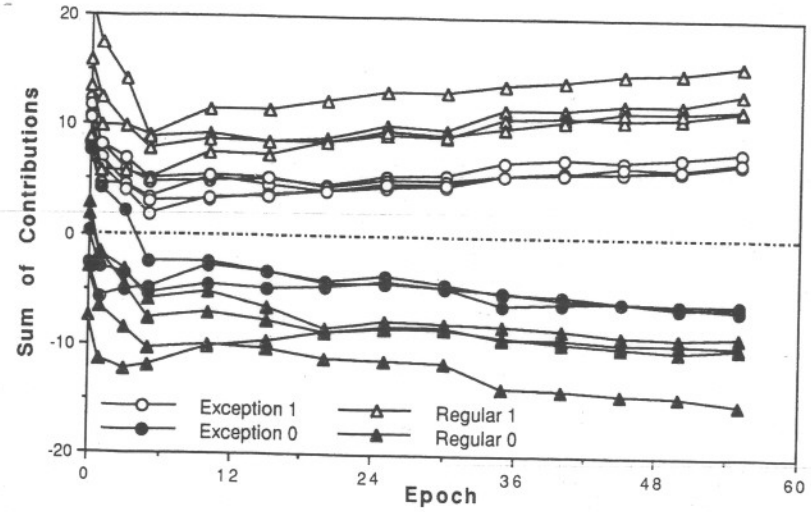

FIG. 11. Relearning curves after damage by hidden unit removal for a rule/exception network.

We have seen that, for sufficiently large-scale networks, the errors caused by damage are primarily on the exceptions, so we generally expect any relearning to simply correct these errors and eventually we achieve perfect performance again. The relearning curves will be like the original learning curves from about epoch 15 in Fig. 5, and so we can expect an enhanced exceptions lost dissociation or a reduced dissociation. There will be no new (rules lost) dissociations created by relearning here.

The only cases where there are large errors on the regular patterns are when the network's performance on everything is rather poor and when we have chance fluctuations causing the errors. If we have all round poor performance, the learning will be like that from a very early epoch in Fig. 5. Small amounts of relearning will enhance any exceptions lost dissociation and larger amounts will eventually correct all the errors. Exactly what happens as we relearn in cases where there are chance fluctuations will depend on details of the numbers involved, but any differential learning rates will always favour the regularities as can be seen occurring in Fig. 11.

It follows that small amounts of relearning may sometimes enhance or create a dissociation with the exceptions lost but never the reverse dissociation. Large amounts of relearning will eventually correct all the errors as long as there remain sufficient computational resources to do so. In cases where the remaining network resources are limited (e.g. very few units or connections remain), it will be the learning of the exceptions that suffers. We conclude that relearning here does not provide an alternative mechanism for creating double dissociations in realistic neural networks.

IMPLICATIONS FOR NEUROPSYCHOLOGY

For two representative problems, we have obtained double dissociations between rules and exceptions in small artificial neural networks. However, we have shown that as we scale up towards larger, more distributed and presumably more realistic networks, the double dissociation disappears, leaving us with a single dissociation in which the most regular mappings are more resilient to damage. We have also seen how this pattern of behaviour can be conveniently understood in terms of learning and damage trajectories in hidden unit activation space.

These basic results are confirmed by more realistic models of reading, spelling and past tense acquisition. Unfortunately, for these more realistic problems, the dissociations will rarely be as clear-cut as we have found in our toy models. There will generally be complex hierarchies of rules, sub-rules and exceptions in the training data that are inevitably confused further by the effects of frequency, noise and differences in training experiences. The conclusion that we can only find a single (regularity-dependent) dissociation in a fully distributed system holds, though finding the appropriate testing examples to demonstrate this clearly will not always be easy.

Our results do not mean that we can never get double dissociation in a realistic neural network system. Obviously, it is easy enough to construct a neural network implementation of a "box and arrow" model and selectively damage it to produce the double dissociations. However, this has no implications for cognitive neuropsychology beyond the implications of the underlying boxes and arrows.

Since our simulations have not covered the whole range of implementational possibilities, our results cannot necessarily imply that it is *never* possible to obtain double dissociation between rules and exceptions in homogeneous neural networks. Rather, this work may be viewed as setting a challenge to modelling researchers to show that rule/exception double dissociations can occur in such networks. We suggest that such a challenge cannot be met; but this conclusion will only be more firmly established (or overturned) by further attempts to meet this challenge. It is to be hoped that, by performing the contribution analysis and checking the damage plots as discussed in this paper, it will be clear in future simulations whether they are sufficiently large scale for the damage results to be reliable. This should help avoid the premature suspicion of the traditional cognitive neuropsychological methodology that has occurred previously (e.g. Chater & Ganis, 1991; Ganis & Chater, 1991; Sartori, 1988; Wood, 1978).

Unfortunately, our results suggest that "sufficiently large scale" will often be "extremely large scale" and beyond our computational resources. Is there anything that can be done about this? Comparing the results obtained for networks of different sizes, we find that it is the global scaling of weights for small networks that gives the best indication of the effects of damage that will be obtained in the corresponding large-scale networks. One reason for this is that global rescaling is a simple deterministic form of damage which has an effect on every part of the network. By contrast, the effect of other kinds of damage, such as the removal of units or connections, will depend crucially on random factors such as the particular units or connections being removed, and since small-scale networks tend not to be fully distributed, these random effects can be quite significant. These latter kinds of random damage effects will be negligible in large-scale, fully distributed networks, in which individual connections or units do not play a crucial role, and the full effect of damage corresponds to the sum of a large number of such contributions which combine to give relatively smooth global effects. According to this analysis, global rescaling gives valuable insights, even when using small networks, since the effects of damage are evenly distributed throughout the whole network in a manner that for large networks offers a good approximation to the other forms of damage.

The simulations reported in this paper address directly the reliability of the double dissociation inference in the context of mappings involving rules and exceptions. What of the double dissociation inference more generally? Our analysis of the logic of double dissociation inference shows that each case should be taken on its merits: it may be that some double dissociation inferences are more reliable than others. Moreover, we have argued that the validity of the inference depends on the class of computational models under consideration. Hence, an important goal for future connectionist neuropsychology is to assess the extent to which connectionist models invalidate or support patterns of neuropsychological inference devised in the context of box and arrow models. But we would urge caution upon connectionist researchers challenging the double dissociation inference in any domain. As our results make clear, results from small-scale cases may not be representative of more realistic networks, and that a systematic analysis such as that conducted here would be necessary to establish that the double dissociation inference is not valid in some other context.

We should also note that in this paper we have been concerned purely with simple feedforward networks. Is it possible that recurrent networks will behave significantly differently? If the recurrent connections are simply supplying additional context information, it seems unlikely that they will make much difference, since we would expect this information to be treated in the same way as additional inputs to a feedforward network. If they are implementing more complicated structures, such as output buffers or basins of attraction, then it will clearly be necessary to carry out a more detailed analysis of each case. The minimum set of checks outlined in this paper can still be applied to ensure that the network is sufficiently large to avoid the small-scale artefacts.

In this connection, an interesting case for future study not covered here is Plaut's observation of a double dissociation between "abstract" words (specifically, words coded by sparse patterns) and "concrete" words (coded by less sparse patterns) in simulations of deep dyslexia (Plaut & Shallice, 1993; Plaut, in press b). Here the dissociation is based on sparseness rather than regularity in the training data and the lesions that result in double dissociation occur at two separate locations within a relatively complicated recurrent network.

Inference from double dissociation to modularity of function, like neuropsychological inference in general, is a difficult matter, as we have discussed above; but our studies suggest that neural network models may not, thankfully, introduce further problems of interpretation, at least in the context of dissociations between rules and exceptions. More generally, we conclude that claims (e.g. Chater & Ganis, 1991; Ganis & Chater, 1991;

Sartori, 1988; Wood, 1978) that connectionist modelling casts doubt upon the inference from double dissociation to modularity of function are as yet unjustified.

REFERENCES

Besner, D., Twilley, L., McCann, R.S., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary? *Psychological Review*, *97*, 432–446.

Brown, G.D.A., Loosemore, R.P.W., & Watson, F.L. (1994). Normal and dyslexic spelling: A connectionist approach. In G.D.A. Brown & N.C. Ellis (Eds), *Handbook of spelling*. Chichester: John Wiley.

Bullinaria, J.A. (1994a). *Representation, learning, generalisation and damage in neural network models of reading aloud.* Technical Report, Edinburgh University.

Bullinaria, J.A. (1994b). Connectionist modelling of spelling. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 78–83. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bullinaria, J.A. (1994c). Internal representations of a connectionist model of reading aloud. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 84–89. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bullinaria, J.A. (1994d). *Learning the past tense of English verbs: Connectionism fights back.* Technical Report, Edinburgh University.

Bullinaria, J.A., & Chater, N. (1993). Double dissociation in artificial neural networks: Implications for neuropsychology. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 283–288. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language*, *21*, 9–20.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from analysis of patterns of impaired performance: The case of single-patient studies. *Brain and Cognition*, *5*, 41–66.

Chater, N., & Ganis, G. (1991). Double dissociation and isolable cognitive processes. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 668–672. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Coltheart, M. (1985). Cognitive neuropsychology and the study of reading. In A. Young (Ed.), *Functions of the right cerebral hemisphere.* London: Academic Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Dietterich, T.G., & Bakiri, G. (1991). Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *Proceedings of the 9th National Conference on Artificial Intelligence.* Anaheim, CA: AIII Press.

Dunn, J.C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

Ellis, A.W. (1987). Intimations of modularity, or, the modularity of mind: Doing cognitive neuropsychology without syndromes. In M. Coltheart, G. Sartori, & R. Job (Eds), *The cognitive neuropsychology of language.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Fahlman, S.E. (1988). Faster-learning variations on back-propagation: An empirical study. In *Proceedings of the 1988 Connectionist Models Summer School.* San Mateo, CA: Morgan Kauffmann.

Fahlman, S.E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D.S. Touretzky (Ed.), *Advances in neural information processing systems 2*, pp. 524–532. San Mateo, CA: Morgan Kauffmann.

Funnell, E. (1983). Phonological processing in reading: New evidence from acquired dyslexia. *British Journal of Psychology*, *74*, 159–180.

Ganis, G., & Chater, N. (1991). Double dissociation in modular systems. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 714–718. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Geschwind, N. (1985). Mechanisms of change after brain lesions. *Annals of the New York Academy of Sciences*, *457*, 1–11.

Glushko, R.J. (1979). The organisation and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 674–691.

Gregory, R.L. (1961). The brain as an engineering problem. In W.H. Thorpe & O.L. Zangwill (Eds), *Current problems in animal behaviour*. Cambridge: Cambridge University Press.

Hinton, G.E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*, 185–234.

Hinton, G.E., & Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructures of cognition*, Vol. 1. Cambridge, MA: Bradford/MIT Press.

Hinton, G.E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Kolen, J.F., & Pollack, J.B. (1991). Back propagation is sensitive to initial conditions. In J.E. Moody, S.J. Hanson, & R.P. Lippman (Eds), *Advances in neural information processing systems 3*, pp. 860–867. San Mateo, CA: Morgan Kauffmann.

Kosslyn, S.M., Flynn, R.A., Amsterdam, J.B., & Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition*, *34*, 203–277.

Kramer, A.K., & Sangiovanni-Vincentelli, A. (1989). Efficient parallel learning algorithms for neural networks. In D.S. Touretzky (Ed.), *Advances in neural information processing systems 1*, pp. 40–48. San Mateo, CA: Morgan Kauffmann.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualisations: Revising the verb learning model. *Cognition*, *40*, 121–157.

Marin, O.S.M., Saffran, E.M., & Schwartz, D.F. (1976). Dissociations of language in aphasia: Implications for normal functions. *Annals of the New York Academy of Sciences*, *280*, 868–884.

McCarthy, R.A., & Warrington, E.K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex*, *22*, 359–380.

Morton, J., & Patterson, K.E. (1980). A new attempt at an interpretation, or, an attempt at a new interpretation. In M. Coltheart, K.E. Patterson, & J.C. Marshall (Eds), *Deep dyslexia.* London: Routledge.

Patterson, K.E., Seidenberg, M.S., & McClelland, J.L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R.G.M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology*. Oxford: Oxford University Press.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D.C. (in press a). Relearning after damage in connectionist networks: Towards a theory of rehabilitation. *Brain and Language.*

Plaut, D.C. (in press b). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology.*

Plaut, D.C., & McClelland, J.L. (1993). Generalisation with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual*

Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Plaut, D.C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tenses of English verbs. In J.L. McClelland & D.E. Rumelhart (Eds), *Parallel distributed processing: Explorations in the microstructures of cognition*, Vol. 2. Cambridge, MA: Bradford/MIT Press.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science*, *1*, 115–138.

Sartori, G. (1988). From neuropsychological data to theory and vice versa. In G. Denes, P. Bisiacchi, C. Semenza, & E. Andrewskv (Eds), *Perspectives in cognitive neuropsychology.* London: Lawrence Erlbaum Associates Ltd.

Seidenberg, M.S. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology*, *5*, 403–426.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Sejnowski, T.J., & Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*, 145–168.

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

Small, S.L. (1991). Focal and diffuse lesions in cognitive models. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 85–90. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, *26*, 608–631.

Teuber, H.L. (1955). Physiological psychology. *Annual Review of Psychology*, *9*, 267–296.

Wood, C.C. (1978). Variations on a theme of Lashley: Lesion experiments on the neural model of Anderson, Silverstein, Ritz & Jones: *Psychological Review*, *85*, 582–591.