Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

Brain and Cognition 57 (2005) 146–150

Brain and Cognition

www.elsevier.com/locate/b&c

Functional localization and double dissociations: The relationship between internal structure and behavior

David A. Medler[a,*], Michael R.W. Dawson[b], Alan Kingstone[c]

[a] Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, United States
[b] Department of Psychology, University of Alberta, Edmonton, Alta., Canada
[c] Department of Psychology, University of British Columbia, Vancouver, BC, Canada

Accepted 12 August 2004

Abstract

Lesioning studies are often used in cognitive neuroscience to make inferences about the architecture of cognition. Recently, computational models have been used to address some of the underlying assumptions—such as modularity and locality—often implicitly used when interpreting lesion data. In this article, we explore the "functional localization" assumption and its role in interpreting lesioning data, especially from double dissociations. The functional localization assumption states that units or subunits within an information processing system become functionally specialized for dealing with specific aspects of the input environment. Networks were trained on one of two problems—an abstract "rules and sub-rules" problem, and a more concrete "logic classification" problem—and then systematically lesioned. Networks were analyzed in terms of their overt behavior, and more importantly, in terms of their internal structure. Performance deficits in both form and magnitude could be directly related to the ablated internal structure of the networks. That is, if an ablated area had little or no functional localization, then little or no behavioral dissociations were observed. If, however, the ablated area had very specific internal structure, then very specific behavioral dissociations were observed. It is important to note, however, that there was not a one-to-one correspondence between internal structure and behavioral dissociations, implying that cognitive neuroscientists must be careful when using lesioning data to theorize about the functional architecture of cognition.

1. Introduction

One of the major goals of cognitive neuroscience is to define the functional architecture of cognition. That is, cognitive neuroscientists are interested in discovering the building blocks of cognition and relating these blocks to the physiology of the brain. To aid in this endeavor, cognitive neuroscientists have an arsenal of different tools at their disposal—from behavioral observations following brain lesions, to imaging the intact brain, to computational modeling of specific cognitive functions.

Initially, inferences about the functional architecture were limited to assigning specific cognitive functions to specific brain regions. Perhaps the best-known examples of this form of inference are the observations and conclusions of Paul Broca and Carl Wernicke. Broca's patients (suffering lesions to the posterior left frontal lobe) had relatively normal comprehension, but could not speak. Wernicke's patients (suffering lesions to the posterior superior temporal lobe) could speak, but had little comprehension. Taken together, the observations of Broca and Wernicke illustrate a classic *double dissociation* (DD).

Farah (1994) coined the term "locality assumption" to describe one possible underlying assumption used by cognitive neuroscientists when interpreting lesion

[*] Corresponding author.
E-mail address: dmedler@mcw.edu (D.A. Medler).

0278-2626/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandc.2004.08.034

data, especially data representing double dissociations. Specifically, the locality assumption relates the functional architecture of cognition to the notion of Fodorian modularity; that is, cognitive tasks can be described in terms of isolable subsystems with *limited interactivity*. This type of assumption is often implicit within the interpretation of patient data. For example, if a patient suffers a lesion to the left frontal operculum and subsequently looses the ability to speak yet maintains language comprehension and motor control (i.e., Broca's aphasia), it is assumed that region of the brain is important for language production. If, however, a patient suffers damage to the posterior superior temporal lobe near the superior temporal gyrus and subsequently loses the ability to understand language and speak in coherent sentences (i.e., Wernicke's aphasia), then it is assumed that region of the brain is important for language comprehension.

The locality assumption maintains that these two subsystems can be damaged independently of each other, and that disrupting one module will not affect the processing in the other module. This is because the *limited interactivity* imposed by the locality assumption precludes the ongoing processing in one module from having a direct and measurable effect on the ongoing processing in the other module. Farah (1994) showed—via a number of interactive models—that this strong form of the locality assumption was false; that is, double dissociations could be observed in fully interactive systems. A distinction remains, however, between the *locality assumption* that describes the connectivity pattern between regions, and the *functional localization assumption* that describes the idea that brain regions may be specialized for particular tasks. Indeed, all of the models that Farah (1994) describes and most previous computational models have relied on some form of functional localization to produce double dissociations. This functional localization is either explicitly assumed, or it is implicitly pre-wired into the network structure.

The question that remains is whether or not behavioral dissociations (specifically, double dissociations) can be related to learned functional localization. Consequently, in these studies, we investigated (1) whether networks will develop functional locality across different tasks and architectures, and (2) if observed behavioral dissociations in lesioned networks can be related to the ablation of such functionally local structure.

2. Method

2.1. Problem type

We trained the networks on two different problem sets. The first problem was Bullinaria and Chater's (1995) "rules and sub-rules" task which captures the essences of the quasi-regularity in grapheme-to-phoneme translation. The data set follows the basic "rule" of straight image translation (i.e., reproducing the input pattern on the output units) and the "sub-rule" of flipping the last three bits whenever the first four bits are '0000' or '1111.' The second problem was Bechtel and Abrahamsen's (1991) "logic problem" data set. This problem set consists of four classes of problem: modus ponens (MP: "If... then"), modus tollens (MT: "If... then), alternative syllogism (AS: "Or"), and disjunctive syllogism (DS "Not both . . . and").

2.2. Network architecture

We trained two different types of feed-forward network architectures; the *value unit* architecture uses a Gaussian activation function, and the *integration device* architecture uses a standard sigmoidal activation function (Dawson & Schopflocher, 1992). The "rules and sub-rules" network had 8 input units, 16 hidden units, and 8 output units. The "logic problem" network had 14 input units, 9 hidden units, and 3 output units. Weights and biases were randomized, and the networks were trained with either a modified (value unit architecture) or the standard (integration device) generalized delta rule.

2.3. Lesioning networks

There are many different approaches that researchers have taken to lesioning PDP networks: for example, adding noise to existing connection weights, cutting specific connections between processing units, and removing entire processing units from the network. It is this latter approach that was used in our experiments. In these simulations, we took our intact PDP network and ablated a single processing unit from it. This was accomplished by forcing the ablated units to always generate an internal activation that was equal to zero, regardless of what stimulus was being presented to the network. In other words, each lesioned network, or "patient," was missing one hidden processing unit, and each lesioned network differed from all others in terms of which hidden unit had been destroyed. By restricting ourselves to the removal of entire processing units from the intact network, we place ourselves in a position to take maximum advantage of our knowledge about the internal structure when interpreting behavioral deficits (i.e., qualitative changes in network outputs). Furthermore, the ablation of individual processing units provides a simple and useful context for asking questions about the functional localization assumption.

2.4. Network analyses

We performed three different analyses on the networks: (i) a qualitative measure of the number and type of errors made, (ii) an analysis of the internal structure of the networks, and (iii) a quantitative measure of the simple structure of the network as measured by activation variance.

Following each lesion, the training patterns were represented to the network and the number of incorrect responses was tabulated by thresholding the response to compel the network into a forced-choice paradigm. From this measure, we can determine both the number of errors made, and the types of errors made (in the case of the logic problem).

The second analysis consisted of extracting the "rules" from the networks by interpreting the activation levels of the hidden units within the networks (Medler, McCaughan, Dawson, & Willson, 1999). Basically, a pattern is presented to the network, and the activation pattern across the hidden units is recorded, much like single cell recording. These activation patterns can then be clustered, and the common input features shared by similar activation patterns can be extracted to produce an interpretation of how each unit solves the problem. In other words, we can analyze the internal structure of an intact network to see if functional localization has been learned.

The third analysis consisted of quantifying the local structure of individual units using an analogy of factor analysis. Researchers use factor analysis to find a set of factors whose loadings are maximally interpretable; one way to accomplish this is to rotate a factor structure until its simple structure is optimized. Simple structure is characterized by a number of data points having high factor loadings—that is, they possess the "feature" that the factor captures—while the remaining data points have zero factor loadings because they do not possess the "feature" captured by the factor. The degree of simple structure can be analytically described by the amount of variance within the factor loadings; factor loadings with high degrees of simple structure will have more variance than factor loadings with little simple structure. This same interpretation can be applied to the activation patterns within the hidden units of a network.

3. Results

3.1. Rules and sub-rules

Twenty different networks were trained on the "rules and sub-rules" patterns. Each network was then systematically lesioned, and the number of errors tabulated. The percentage of errors for the "rules" set ranged from 0 to 51%, while the range for the "sub-rules" set was 0–100%. To assess whether double dissociations were observed within the networks, the percentage of errors from the rules were subtracted from the sub-rules; this gives four possible measures of difference: zero—lesions produced no observable behavioral deficits in either group; equal—lesions produced equivalent error percentages in both groups (±10%); positive—lesions produced better performance in the rules group; negative—lesions produced better performance in the sub-rules group. This difference measure was computed for each "patient" for each of the 20 networks. The means (and standard deviations) for the difference groups are as follows: zero = 2.55 (1.58); equal = 4.75 (2.40); positive = 4.75 (1.65); negative = 3.90 (1.86). As can be seen, a behavioral double dissociation is observed: some lesioned networks do better on the rules, and some lesioned networks perform better on the sub-rules.

Table 1 shows a detailed breakdown of the types of errors made for a single network. An analysis of the internal structure of the network showed that units H1, H2, H3, H4, H5, H6, H9, H10, H14, and H15 all encoded single bits within the input pattern. Consequently, when these units were ablated, performance on the "rule" patterns was affected. Conversely, units H7, H8, and H12, encoded the specific "sub-rule" patterns, and when lesioned, produced deficits on the sub-rule patterns. Finally, units H11, H13, and H16 had no interpretable structure and produced no substantial behavioral deficits when lesioned. When we compare the number of errors with the simple structure, we produce a significant correlation. ($R^2 = 0.78$, $p < .0001$), indicating that the network had developed functional localization.

3.2. The logic problem

For brevity, we only report the results of one value unit network and one integration device network that have been trained on the logic problem. Table 1 shows the total number of errors made for each problem type, and the total number (56 maximum) of error types (Err). For the value unit architecture, the average number of error types was 5.1 (2.83), while the average number of total errors summed across problem types was 120.7 (84.7). There is a strong positive linear correlation between the number of error types and total number of errors ($R^2 = 0.8034$, $p < .005$). The integration device network, on the other hand, produced an average of 9.67 (2.73) error types and an average total of 194.4 (90.0) errors. Both of these values were significantly more than the value unit network ($t(14) = 10.69$, $p < .001$; $t(14) = 4.66$, $p < .01$, respectively). Although positive, the correlation between the number of error types and total number of errors for the integration de-

Table 1
Percent difference errors for the "rules and sub-rules" network and the error types and total number of errors made for the "logic problem" network

Architecture		Hidden unit ablated								
Rules and sub-rules		H1	H2	H3	H4	H5	H6	H7	H8	
Value unit	Diff	−25	−42.9	−37.5	−25.0	−35.7	−25.0	50	31.3	
		H9	H10	H11	H12	H13	H14	H15	H16	
	Diff	−25.9	0	0	43.7	0	−25	−28.6	−3.6	
Logic problem		H1	H2	H3	H4	H5	H6	H7	H8	H9
Value unit	AS	96	48	66	67	0	0	1	0	9
	DS	0	48	59	0	0	192	0	0	18
	MP	24	48	9	48	96	0	0	0	0
	MT	72	0	58	21	96	0	1	0	9
	Err	7	4	9	6	8	6	2	0	4
Integration	AS	28	167	22	13	54	48	105	0	27
	DS	66	0	66	0	30	67	48	144	98
	MP	14	54	39	0	70	24	72	67	44
	MT	43	96	21	9	25	48	24	96	21
	Err	10	12	9	4	14	8	10	9	11

vice network merely approached significance ($R^2 = 0.3365$, $p = .10$).

Importantly, behavioral analyses of the networks show a double dissociation for both the value unit network and the integration device network. Specifically, the value unit network shows a double dissociation between H5 (cannot perform "If ... then" problems) and H6 (fails on "Not both ... and" problems). Similarly, the integration device network shows a behavioral double dissociation between units H2 (is able to solve "Not both ... and" problems) and H8 (is perfect on "Or" problems); it should be noted, however, that both these units also have difficulties with the "If ... then" problems.

Analyses of the internal structure of the value unit network confirmed that H6 specialized in detecting the connective "Not both ... and." Conversely, analysis of the internal structure of H5 showed that it was tuned to the connectives "If... then" and "Not both ... and." In other words, the functional localization of H5 is more than the behavioral analysis would indicate. It should be noted that the analysis of unit H8 showed no internal structure, and when ablated, this produced no behavioral deficits. For the integration device network, analyses of the units confirmed that H2 detected the connectives "If ... then" and "Or" while H8 was specialized for processing the connectives "If ... then" and "Not both ... and." Thus, it appears that local behavioral deficits are related to the ablation of local internal structure, regardless of network architecture.

Finally, our third analysis technique revealed a significant positive linear relationship between the simple structure of the units—as measured by variance of the unit activations—and the number of error types ($R^2 = 0.5815$, $p < .05$) and the total number of errors ($R^2 = 0.8074$, $p < .005$) for the value unit networks. Although positive, the correlations for the integration device networks approached significance for the number of error types ($R^2 = 0.4072$, $p = .06$), but not for the total number of errors ($R^2 = 0.1307$, $p = \text{ns}$).

4. Discussion

It was found that local lesions produce very local and severe impairments in different network architectures, and on different types of problems. This was confirmed by a behavioral analysis, by an analysis of the internal structure of the networks, and by a quantitative measure of the simple structure of the networks. In other words, for local lesions to produce local behavioral deficits, some form of functional localization must be present. Therefore, these (and previous) models confirm that the strong form of the locality assumption is false, but these models indicate that for a double dissociation to occur within a computational model, the model must have some form of functional localization.

More importantly, however, these results have a strong implication for cognitive neuroscientists studying the localization of function via behavioral and lesion data. In our study, we found that the functionally local structure of the network as indicated by the internal analysis was not necessarily the same as the local structure implied by the behavioral performance of the network. The network structure showed more of a coarse coding organization than a strict modular organization; that is, multiple processing areas contribute to the solving of any particular problem. Thus, based on the behavioral data alone, our conclusions about the internal structure would have been incorrect. Therefore, this implies that cognitive neuroscientists must be very careful when assigning functions to local structure based on behavioral data alone.

References

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Cambridge, MA: Blackwell.

Bullinaria, J. A., & Chater, N. (1995). Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes, 10*, 227–264.

Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks of non-monotonic processors for pattern classification. *Connection Science, 4*, 19–31.

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the locality assumption. *Behavioral and Brain Sciences, 17*, 43–104.

Medler, D. A., McCaughan, D. B., Dawson, M. R. W., & Willson, L. (1999). When local isn't enough: Extracting distributed rules from networks. In: *Proceedings of the 1999 international joint conference on neural networks* (pp. 305i–305vi). Washington DC.