

One word, one module?

Patrick Juola

Duquesne University
Pittsburgh, PA 15282 USA
juola@mathcs.duq.edu

In a discussion of (Juola and Plunkett, 2000), (Thomas and de Wet, 1998) argue that the apparent double dissociations found in that study simply show that the backpropagation network used did not have enough hidden units and that the network did (or could) not develop “fully distributed internal representations.” In other words, the dissociations found were, in their opinion, probably not just pathological cases, but represented hidden modularity that had somehow sneaked into the system despite the original researchers’ best efforts at prevention. The logic of inferring functional modularity from dissociations remains strong and compelling, even if (as will be argued) wrong.

In addition to the points raised by Dunn and Kirsner (this issue) about interpreting dissociations, there are two additional points of critical importance. The first is simply one of philosophical plausibility; are there reasons beyond simple dissociability to assume that two tasks are structurally differentiable? What is the *function* of the (separated) function? Is the separation itself reasonable or useful? The second is an apparent assumption that a cognitive module is either destroyed or undamaged, but that “damaged” is not a state to be further investigated. Superficially, this makes sense, as a module is (presumably) composed of neurons, and in a brain lesion, the missing neurons lose their function entirely. However, unless each module consists of exactly one neuron, a lesion is likely to affect some, but not all, of a module, and hence *alter* the performance of that module, perhaps systematically, perhaps randomly, perhaps even for the better (Juola and Plunkett, 2000; Thomas and Karmiloff-Smith, *ress*). In a more realistic case, neurons may not only cease functioning, but may just behave differently, introducing noise into the modules and again altering behavior, but not destroying it.

A simple investigation and crude model of aphasia (specifically, anomia) can illustrate this. The relevant psychological facts are fairly accessible (Goodglass and Wingfield, 1997); upon brain damage, some patients show an inability to name objects when confronted with them, while being able to select them from a list or be phonologically prompted (e.g., with the initial letter). Many actually show category-specific

anomia, being able, to name (e.g.) body parts but not animals, carpenter’s tools but not vegetables. Following conventional theory, we assume an inability to retrieve the phonology of the word from its semantic representation, and constructed a connectionist network on that basis.

The network was of a somewhat unusual, crude, and psychologically *implausible* type — a Hopfield network (Hopfield, 1982; Hertz et al., 1991). Unlike typical multilayer perceptions, a Hopfield network contains only one layer and does not “learn.” Instead, it implements a form of content-addressable memory where an initial pattern is placed in the single layer, and then the pattern changes to match one of the stable memorized states. Connections exist between every pair of units, and their weights are not learned, but are instead defined by the network’s creator via simple arithmetic regarding the patterns to be memorized. A network was constructed to use ten random English monosyllables (coded into 51 phonological units via the PGPfone coding (Juola and Zimmermann, 1996), and an equal number of random binary activations representing the semantics of the monosyllables). Upon network construction, the network was capable of recreating/accessing the phonology of seven of the ten “learned” words from semantics alone.

This network, in turn, was “lesioned” 10,000 times by deleting approximately 10% of the connections. As expected, average performance of the damaged network was lower (acquired anomia), with different lesions resulting in different words lost. Over those 10,000 lesions, however, every possible double dissociation between individual words was observed. In other words, within this model’s vocabulary, each and every word apparently possesses its own individual “module” or “function.”

In light of the assumptions outlined above, this is *exactly* the wrong analysis, and even the wrong level of analysis. Unlike a multilayer perceptron, every unit is simultaneously an input and an output, and there are no hidden units to recruit to spot features and categories of the input patterns. Connections exist only between pairs of units, and the values of these connections are related only to the correlation between the individual activations levels across patterns. The

representation cannot be less than “fully distributed.” And yet, the network dissociates — doubly.

There are three possibilities, none pleasant. First, if one regards each connection and every unit as an individual module, then any damage will, of course, involve separate modules. This, however, implies that the brain, and by extension cognition, can only be understood on a basis of individual neurons and synapses. A second possibility is to complain that the gremlins of localist internal representations of patterns have somehow slipped into a network without internal representations and that does not represent patterns, an unlikely scenario. The third is to treat the system as a single module, but to recognize that random damage to the system as a whole will result in altered and randomized outputs. Thus, the appearance of double dissociations in this network is likely to be the meaningless result of chance.

What, then, are meaningful dissociations in anomia? (Goodglass and Wingfield, 1997)[p. 23, cites omitted] provides an example : “[Categories] that *typically* have been impaired in patients *with postencephalitic lesions of the inferior temporal lobe* have been animals, fruits, and vegetables, and they have been contrasted with manipulable objects, such as tools and utensils, which are usually spared in these patients” (emphasis mine). By observing localizations in the brain that correspond to particular categories, and further noting that these categories are lost more often and more uniformly than chance predicts, it is suggestive, but not conclusive, that naming fruits involves particular mental functions, and that a specific lobe is involved in representing fruits. Paraphrasing (Thomas and de Wet, 1998), any lesioning will produce a distribution of deficits. A functional model of the human mind should be evaluated not on the possibility of producing a single given deficit, but on the probability of producing the appropriate distribution.

References

- Goodglass, H. and Wingfield, A. (1997). Word finding deficits in aphasia. In Goodglass, H. and Wingfield, A., editors, *Anomia : Neuroanatomical and Cognitive Correlates*, pages 3–30. Academic Press, San Diego.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Cognition*. Addison-Wesley, Reading, MA.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences, USA*, volume 79, pages 2554–2558.
- Juola, P. and Plunkett, K. (2000). Why double dissociations don’t mean much. In Cohen, G., Johnston, R. A., and Plunkett, K., editors, *Exploring Cognition : Damaged Brains and Neural Networks*. Psychology Press/Taylor and Francis, Philadelphia, PA.
- Juola, P. and Zimmermann, P. (1996). Whole-word phonetic distances and the PGPfone alphabet. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-96)*, Philadelphia, PA.
- Thomas, M. and Karmiloff-Smith, A. (in press). Are developmental disorders like cases of adult brain damage? implications from connectionist modelling. *Behavioral and Brain Sciences*, 26.
- Thomas, M. S. C. and de Wet, N. M. (1998). Stochastic double dissociations in distributed models of semantic memory. In *5th Neural Computation and Psychology Workshop : Connectionist Models in Cognitive Neuroscience*, Birmingham, UK.