Lesioned Attractor Networks as Models of Neuropsychological Deficits

David C. Plaut
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213–3890

RUNNING HEAD: Lesioned Attractor Networks

Correspondence:

David C. Plaut
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213–3890
Phone: (412) 268–5145
Fax: (412) 268–5060
Email: plaut@cmu.edu

1 Introduction

A highly controversial issue concerning the neural implementation of cognition is the degree to which cognitive functions are localized to particular brain regions vs. distributed throughout large areas of cortex. Connectionist models using distributed representations provide a natural formalism for expressing how cognitive processes can be distributed over large numbers of neuron-like processing units (see the Overview article on "Cognitive Modeling: Psychology and Connectionism"). In fact, many characteristics of these models echo back to Lashely's original ideas about mass action and equipotentiality. However, the possibility that cognitive functions are distributed widely in the cortex is seriously challenged by the remarkably selective cognitive deficits than can occur in brain-injured patients (see Shallice, 1988, for a review). Can the nature of distributed computation in networks be reconciled with these selective deficits? In fact, as the work described below illustrates, connectionist networks are leading to new insights about how disorders of brain function can give rise to disorders of cognition, challenging traditional assumptions about the modular organization of cognitive functions (see Farah, in press, for discussion).

2 Attractors and Damage

Researchers have begun to explore the degree to which the effects of damage in connectionist models of cognitive processes reproduce certain types of neuropsychological deficits. Many of these investigations use recurrent, interactive networks that develop attractors for familiar patterns of activity (see the Overview article on "Computing with Attractors"). In an attractor network, units interact in such a way that the initial activity pattern generated by

an input gradually settles to the nearest attractor pattern. If the state of each unit is represented along a separate dimension in a multi-dimensional *state space*, then each attractor corresponds to a particular point within this space, and the set of patterns that settle to it corresponds to a region around it called its *basin* of attraction. Although recurrent networks can also exhibit more complex dynamic behavior, such as "limit cycle" and "chaotic" attractors, virtually all existing attempts to model cognitive and neuropsychological phenomena have relied on "point" attractors.

Lesioning a network, by removing some proportion of its units or connections, alters the settling behavior of the remaining units. In state space, the effects of damage amount to distortions of the shapes and positions of attractor basins. As a result, the initial pattern of activity for an input now may fall within a neighboring basin, giving rise to an incorrect response. The detailed pattern of correct and incorrect performance will depend on specifics of the layout of attractor basins (as a function of the task, network architecture, and training procedure) and how the basins are distorted by the lesion (as a function of the location, severity, and type of damage).

3 Models of Neuropsychological Deficits

Brain damage can produce selective impairments in a wide range of cognitive domains, including high-level vision, attention, written and spoken language, learning and memory, planning, and motor control (Shallice, 1988). In many of these domains, lesioning a connectionist model of the normal process leads to analogous deficits, suggesting that the models capture important properties of the normal cognitive system.

3.1 Deep Dyslexia

The class of neuropsychological impairments which have received perhaps the greatest theoretical attention are those that involve word reading, the so-called acquired dyslexias. Of these, *deep dyslexia* is among the most perplexing (Coltheart, Patterson, & Marshall, 1980). The hallmark characteristic of the reading behavior of deep dyslexic patients is the occurrence of *semantic* errors, such as reading the word RIVER as "ocean" or DARK as "night." These patients also make *visual* errors (e.g. SCANDAL => "sandals"), suggesting a second impairment. However, if two separate lesions are involved, why do visual errors virtually always co-occur with semantic errors?

Hinton and Shallice (1991) demonstrated that the co-occurrence of visual and semantic errors is a natural consequence of a single lesion to an attractor network trained to derive the meanings of written words. They trained a recurrent back-propagation network to map from the visual form (orthography) of 40 three- or four-letter words to a simplified representation of their semantics, described in terms of 68 predetermined semantic features (e.g. *brown*, *made-of-wood*, *for-cooking*). After training, lesions throughout the network resulted in both semantic errors (e.g. CAT => "dog") and visual errors (e.g. CAT => "cot"), similar to those observed in deep dyslexia. Essentially, in order for the network to solve the task, the layout of attractor basins must be sensitive to both visual and semantic similarity; as a result, these metrics are reflected in the types of errors that occur as a result of damage (see Figure 1).

Insert Figure 1 about here.

More recently, Plaut and Shallice (1993a) have extended these initial findings in two ways. First, they established the generality of the co-occurrence of visual and semantic errors by showing that it does not depend on particular characteristics of the network architecture, the training procedure, or the way responses are generated from semantic activity. Second, they extended the approach to account for many of the remaining characteristics of deep dyslexia, including the effect of concreteness on reading accuracy (e.g. TABLE is more likely to be read correctly than TRUTH) and its interaction with visual errors; the occurrence of other types of errors (e.g. *visual-then-semantic*; SYMPATHY => "orchestra," presumably via *symphony*); greater confidence in visual than in semantic errors; relatively preserved lexical decision; and the existence of different subvarieties of deep dyslexia. The only major additional assumption in these extensions is the hypothesis that abstract words have fewer semantic features than concrete words, which causes the network to form weaker attractors for them. The same general approach has also been used to account for the perseverative and semantic influences on the errors that optic aphasic patients make in naming visually presented objects (Plaut & Shallice, 1993b), and the degree of recovery and generalization in cognitive rehabilitation studies with acquired dyslexic patients (Plaut, in press).

3.2 Neglect Dyslexia

Mozer and Behrmann (1990) have accounted for another reading disorder, known as *neglect dyslexia*, based on principles very similar to those used in the deep dyslexia simulations. Patients with neglect dyslexia often ignore the contralesional (typically left) portion of written material, even when it falls entirely within the intact portions of their visual fields (Riddoch, 1991). Incorrect responses to letter strings typically consist of letter omissions (e.g. CHAIR

=> "hair"), substitutions (e.g. HOUSE => "mouse"), or additions (e.g. LOVE => "glove"). The severity of the deficit is influenced by both peripheral and central manipulations. Thus, the accuracy of reading a letter string is better when the stimulus is presented further to the right, or when it forms a word.

In the model that Mozer and Behrmann used (Mozer, 1991), bottom-up letter information interacts with top-down lexical/semantic knowledge to form attractors for words. The bottom-up input is gated by an Attentional Mechanism (AM) that forms a spatially contiguous "spotlight" on the basis of where letter features occur. Letter features that fall outside the spotlight are much less likely to be transmitted to the word recognition system. Mozer and Behrmann model the attentional impairment in neglect dyslexia by introducing a monotonic gradient of damage to the connections from the letter features to the AM, with damage most severe on the left. This damage biases the AM towards forming a spotlight that includes only the rightmost letters of an input string, resulting in corrupted letter input to the word recognition system. The lexical attractors can often reconstruct the correct pattern of activity, particularly when the entire input forms a word. However, when this process fails, the result is often another word that differs from the presented word only on the left. Reading accuracy is better if the letter string is presented further to the right because the damage from these positions to the AM is less severe.

3.3 Hemispatial Neglect

In traditional cognitive neuropsychological accounts, investigators have often stipulated the existence of a specialized module in the cognitive system whenever a brain-damaged patient exhibits a selective deficit in some specific aspect of cognitive function. For example, patients

with *hemispatial neglect*—a more general attentional deficit than neglect dyslexia—are abnormally slow to shift their attention from a pre-cued ipsilesional location to a contralesional stimulus. This has been interpreted in terms of damage to a specific "disengage" module (Posner, Walker, Friedrich, & Rafal, 1984). However, Cohen, Romero, Servan-Schreiber, and Farah (in press) have reproduced this deficit in shifting attention by unilaterally damaging a connectionist model with no special disengage module. Rather, in the model, attention is allocated based on competitive interactions among units representing different spatial locations. The unilateral damage causes an imbalance in this competition, making it difficult for attention to be captured by contralesional locations.

3.4 Prosopagnosia

Another example of the influence of connectionist models on the interpretation of neuropsychological deficits concerns patients with "prosopagnosia." These patients fail to name familiar faces and have no conscious recollection of them, and yet often show evidence of recognition on tasks, such as priming or name relearning, that measure covert knowledge. This behavioral dissociation has led some researchers to propose separate mechanisms of overt and covert recognition. But Farah, O'Reilly, and Vecera (1993) have demonstrated that this dissociation between overt and covert face recognition can arise in a single system. Partial damage can virtually eliminate overt naming performance while still leaving sufficient residual information available to support above-chance performance on more indirect tests of semantic knowledge.

3.5 Category-Specific Semantic Deficits

As a final example, connectionist models are contributing to the question of whether knowledge is organized by modality or by category. Apparent evidence for category specificity comes from the finding that some brain-injured patients are selectively impaired in recognizing and recalling information about animate objects (e.g. animals) vs. inanimate objects (e.g. tools). Warrington and Shallice (1984) suggested that these deficits could be accounted for if semantics were instead organized by modality, under the dual assumption that visual semantics is impaired and that animate objects rely more heavily on visual than on functional semantics. This explains why the patients were also impaired on inanimate categories, such as gemstones and fabrics, that rely primarily on visual attributes. However, in the absence of a computational model, this account was rejected in favor of categorical organization when it was noted that patients had difficulty both with functional as well as visual aspects of animate objects. But Farah and McClelland (1991) revived the original hypothesis by accounting for the entire pattern of deficits in an interactive, distributed model in which visual and functional semantics interact. The simulation accounts for the patients' paradoxical impairment in recalling functional information about animate objects because functional semantics normally relies on interactions with intact visual semantics to settle to the correct pattern.

4 Discussion

Connectionist networks would appear *a priori* to be an appropriate formalism within which to develop computational models of neuropsychological disorders. Although the specific relationship between these networks and neurobiology is far from clear, the belief that rep-

resentation and computation in these networks resembles neural computation at some level remains one of their strongest attractions. As the research reviewed here illustrates, the computational formalism of attractors, and their behavior under damage, can lead to a deeper understanding of a wide range of neuropsychological phenomena. However, most simulations to date have addressed impairments only within very specific processing domains. A central challenge for future work is to extend these preliminary findings in developing more comprehensive simulations of broader aspects of normal and impaired cognitive processing. Nonetheless, even at this early stage of research, the finding that the behavior of attractor networks after damage resembles that of neurological patients supports the claim that the apparent similarity of artificial and biological neural networks is, in fact, substantive.

REFERENCES

Cohen, J. D., Romero, R. D., Servan-Schreiber, D., and Farah, M. J., in press, Disengaging from the disengage function: The relation of macrostructure to microstructure in parietal attentional deficits. Journal of Cognitive Neuroscience.

Coltheart, M., Patterson, K. E., and Marshall, J. C., Eds., 1980, Deep Dyslexia. London: Routledge and Kegan Paul.

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

Farah, M. J. and McClelland, J. L., 1991, A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. Journal of Experimental Psychology: General, 120:339–357.

Farah, M. J., O'Reilly, R. C., and Vecera, S. P., 1993, Dissociated overt and covert recognition as an emergent property of a lesioned neural network. Psychological Review, 100:571–588.

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

Mozer, M. C., 1991, The Perception of Multiple Objects: A Connectionist Approach. Cambridge, MA: MIT Press.

Mozer, M. C. and Behrmann, M., 1990, On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. Journal of Cognitive Neuroscience, 2:96–123.

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

- Plaut, D. C. and Shallice, T., 1993a, Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology, 10:377–500.

Plaut, D. C. and Shallice, T., 1993b, Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. Journal of Cognitive Neuroscience, 5:89–117.

Posner, M. I., Walker, J. A., Friedrich, F. J., and Rafal, R. D., 1984, Effects of parietal injury on covert orienting of visual attention. Journal of Neuroscience, 4:1863–1874.

Riddoch, M. J., Ed., 1991, Cognitive Neuropsychology: Neglect and the Peripheral Dyslexias. Hillsdale, NJ: Lawrence Erlbaum Associates.

- Shallice, T., 1988, From Neuropsychology to Mental Structure. Cambridge: Cambridge University Press.

Warrington, E. K. and Shallice, T., 1984, Category specific semantic impairments. Brain, 107:829–853.

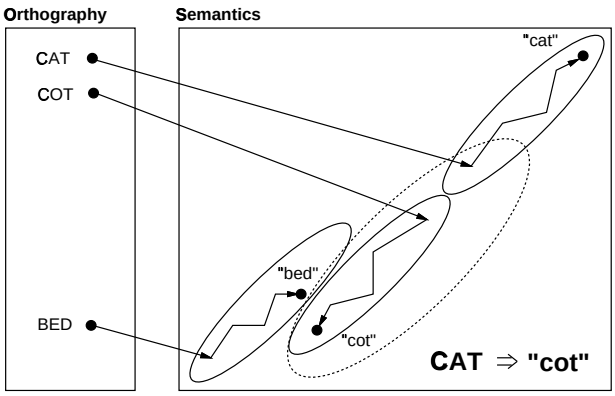
Orthography
Semantics
CAT
COT
BED
"cat"
"bed"
"cot"
CAT ⇒ "cot"

Figure captions

Figure 1: How damage to semantic attractors can cause visual errors. The solid ovals depict the normal basins of attraction; the dotted one depicts a basin after semantic damage. (Reprinted from Plaut & Shallice, 1993a, p. 393)