Simulating Brain Damage

Adults with brain damage make some bizarre errors when reading words. If a network of simulated neurons is trained to read and then is damaged, it produces strikingly similar behavior

by Geoffrey E. Hinton, David C. Plaut and Tim Shallice

In 1944 a young soldier suffered a bullet wound to the head. He survived the war with a strange disability: although he could read and comprehend some words with ease, many others gave him trouble. He read the word *antique* as "vase" and *uncle* as "nephew."

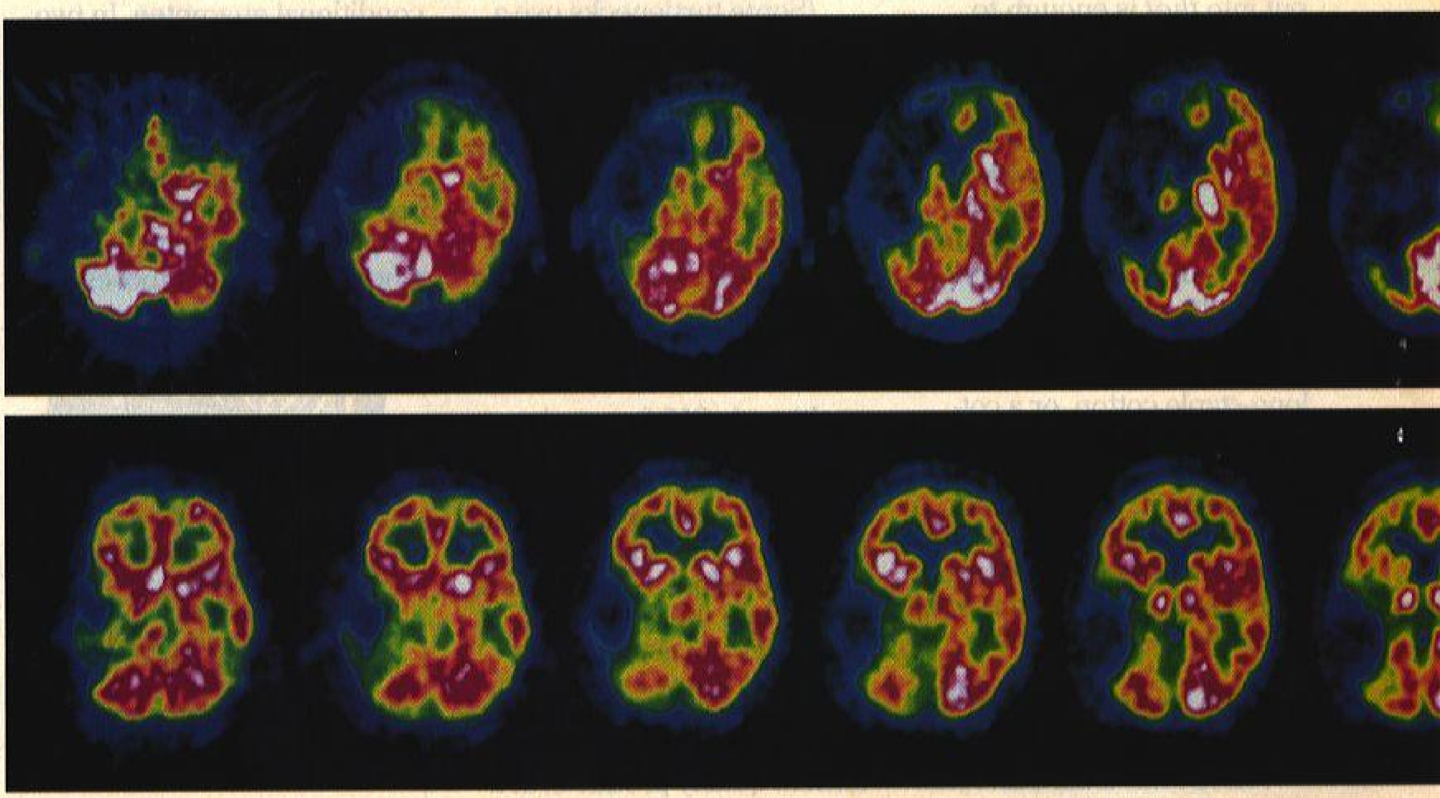
The injury was devastating to the patient, G.R., but it provided invaluable information to researchers investigating the mechanisms by which the brain comprehends written language. A properly functioning system for converting letters on a page to spoken sounds reveals little of its inner structure, but when that system is disrupted, the peculiar pattern of the resulting dysfunction may offer essential clues to the original, undamaged architecture.

During the past few years, computer simulations of brain function have advanced to the point where they can be used to model information-processing pathways. We have found that deliberate damage to artificial systems can mimic the symptoms displayed by people who have sustained brain injury. Indeed, building a model that makes the same errors as brain-injured people do gives us confidence that we are on the right track in trying to understand how the brain works.

We have yet to make computer models that exhibit even a tiny fraction of the capabilities of the human brain. Nevertheless, our results so far have produced unexpected insights into the way the brain transforms a string of letter shapes into the meaning of a word.

When John C. Marshall and Freda Newcombe of the University of Oxford analyzed G.R.'s residual problems in 1966, they found a highly idiosyncratic pattern of reading deficits. In addition to his many semantic errors, G.R. made visual ones, reading *stock* as "shock" and *crowd* as "crown." Many of his misreadings resembled the correct word in both form and meaning; for example, he saw *wise* and said "wisdom."

Detailed testing showed that G.R. could read concrete words, such as *table*, much more easily than abstract words, such as *truth*. He was fair at reading nouns (46 percent correct), worse at adjectives (16 percent), still worse at verbs (6 percent) and worst of all at



BRAIN IMAGES show damage to the language-processing areas of patients with acquired dyslexia, which can now be modeled by artificial neural networks. (These positron-emission tomography scans, made by Cathy J. Price and her colleagues at the MRC Cyclotron Unit in London, measure activity of the brain in successive horizontal slices, starting at the top. Low

function words, such as *of* (2 percent). Finally, he found it impossible to read wordlike nonsense letter strings, such as *mave* or *nust*.

Since then, clinicians have studied more than 50 other patients who make semantic errors in reading aloud, and virtually all of them show the same strange combination of symptoms. In 1973 Marshall and Newcombe described two contrasting types of acquired dyslexia. So-called surface dyslexics misread words that are pronounced in an unusual way, often giving the more regular pronunciation; a surface dyslexic might read *yacht* as "yatched." In contrast, a "deep" dyslexic patient like G.R. might read *yacht* as "boat."

To explain the existence of these two types of dyslexia, Marshall and Newcombe proposed that the information processed in normal reading travels along two distinct, complementary routes. Surface dyslexics retain the phonological route, which relies on common spelling-to-sound correspondences. Deep dyslexics, meanwhile, retain the semantic route, which allows the meaning of a word to be derived directly from its visual form (when it can be derived at all). A person reading words aloud via the semantic route derives pronunciation entirely from meaning.

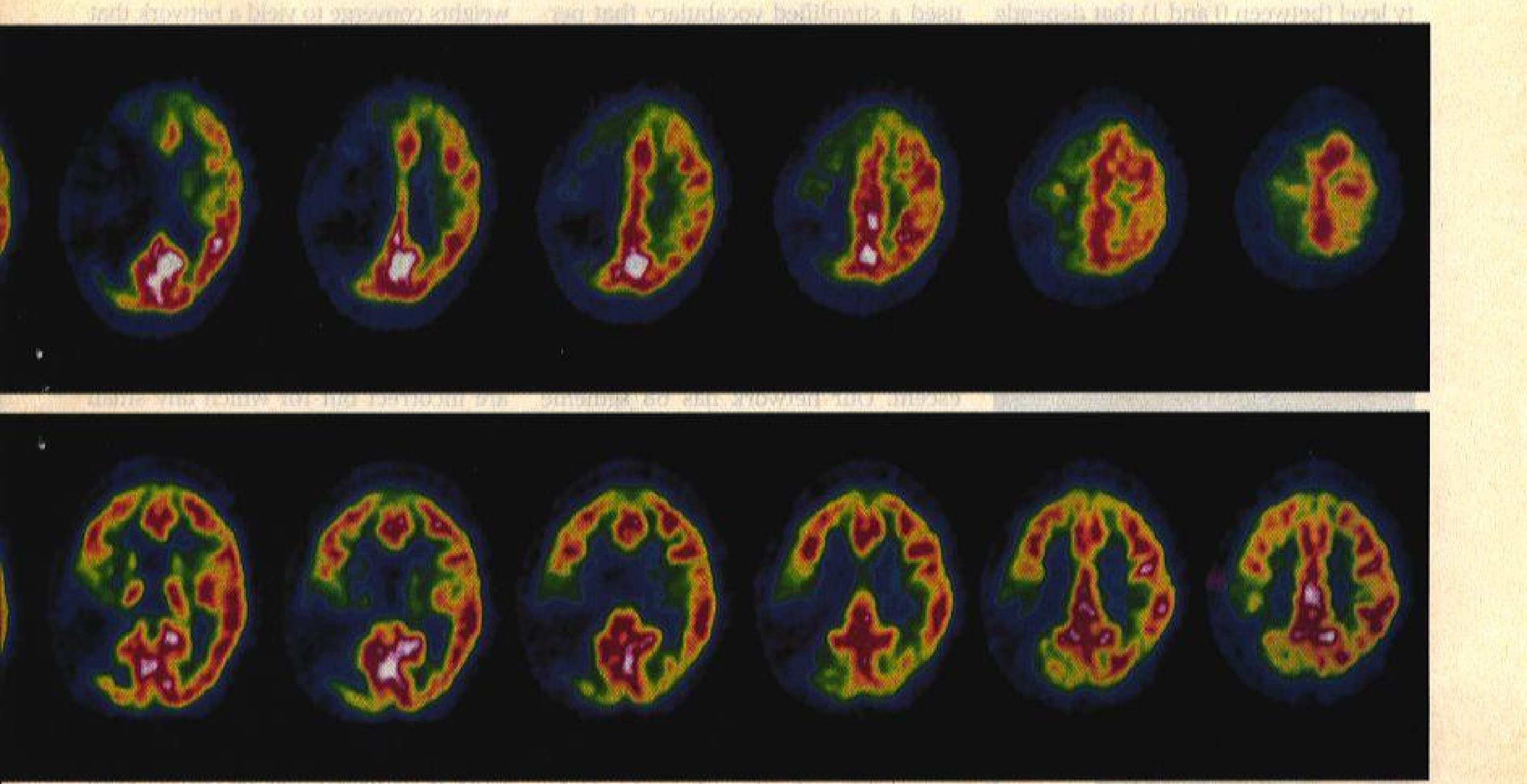
According to Marshall and Newcombe, the errors produced by deep dyslexics reflect how the semantic route operates in isolation. Later empirical findings suggest that this account is oversimplified, but the notion of a semantic route is still generally accepted. It now seems likely that deep dyslexics not only lose their phonological route but have damage somewhere along the semantic one as well.

The hypothesis that reading depends on multiple routes that can be separately damaged has proved fruitful in classifying patients but less useful in understanding the precise nature of their injuries. Max Coltheart of Macquarie University in Australia and Eleanor M. Saffran of Temple University have both proposed, for example, that the reading of deep dyslexics may bear a strong resemblance to that of patients who have only the right hemisphere of their brain functioning.

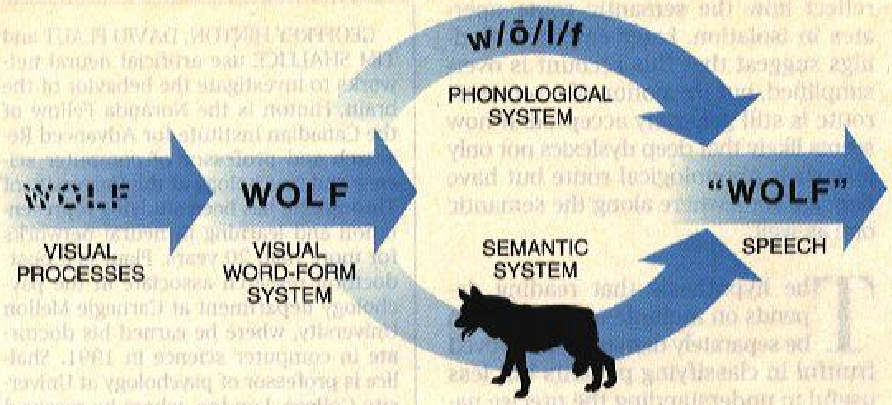
This explanation, however, provides little insight into the highly characteristic pattern of errors that typically occurs in acquired dyslexia. Any detailed explanation of how errors arise and why they form consistent patterns requires a model of how that information is processed in each route—and of how this processing goes wrong when the neural circuitry is damaged. Psychologists often use abstract, algorithmic descriptions of the way that the brain handles information. These descriptions obviously cannot be subjected to the kinds of injuries that brain cells may incur.

As a result, we have turned to neural networks—idealized computer simulations of ensembles of neurons. We have developed networks that perform the role of the semantic route, and then we have selectively removed connections between neurons to see how their be-

GEOFFREY HINTON, DAVID PLAUT and TIM SHALLICE use artificial neural networks to investigate the behavior of the brain. Hinton is the Noranda Fellow of the Canadian Institute for Advanced Research and professor of computer science and psychology at the University of Toronto. He has been studying representation and learning in neural networks for more than 20 years. Plaut is a postdoctoral research associate in the psychology department at Carnegie Mellon University, where he earned his doctorate in computer science in 1991. Shallice is professor of psychology at University College, London, where he received his doctorate in 1965. His research has mainly focused on what can be understood about the normal cognitive system by studying impairments resulting from neurological disease, with occasional forays into the organization of short-term memory and the nature of will and consciousness.



levels of activity appear in blue and high levels in white.) One patient (*top row*) has lost almost all function in the left hemisphere of the cerebral cortex, except for the most posterior regions. The other has sustained damage to the parietal and temporal lobes of the left hemisphere, regions generally believed to be crucial for processing language.



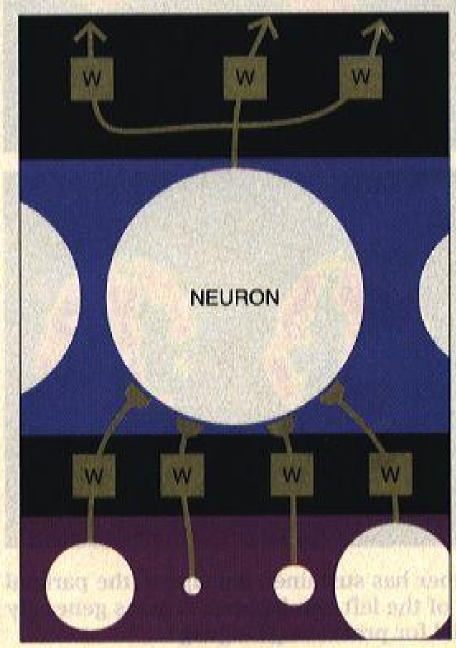
TWO PATHWAYS in the brain are responsible for the mental processing and pronunciation of written words. One (the phonological route) derives pronunciation from spelling, the other (the semantic route) from meaning. Deep dyslexics have lost the phonological route completely and have suffered damage to the semantic route as well.

havior changes. A few years ago we designed a simple network to mimic the semantic route and found that damaging any part of it could reproduce several of the symptoms of deep dyslexia. We have since made more detailed models to learn which aspects of neural-network architectures were responsible for this behavior. We have also extended the approach to account for additional symptoms of deep dyslexia.

Our models of the semantic route consist of interconnected units representing neurons. Each neuron unit has an activity level (between 0 and 1) that depends on the inputs it receives from other neurons. Connections between units have an adjustable weight that specifies the extent to which the output of one unit will be reflected in the activity of the unit it is feeding. These weights, along with the pattern of connections among neurons, determine the computation that the network performs.

The first version of our network consisted of three sets of units: "grapheme" units, each of which represented a particular letter in a specific position within the word; "sememe" units that represented the meanings of words; and a layer of intermediate units that make it possible to learn complex associations. A completely general network would require 26 grapheme units for each position within a word, but we used a simplified vocabulary that permitted a smaller number. The grapheme units in the first position were all consonants, for instance, and those in the second, all vowels.

The sememe units do not correspond directly to individual word meanings but rather to semantic features that describe the thing in question. The word *cat* activates such units as "mammal," "has legs," "soft" and "fierce." Units representing such semantic features as "transparent," "tastes strong," "part of limb" or "made of wood" remain quiescent. Our network has 68 sememe units representing both physical and functional attributes of a word's definition. Each word that we chose was represented by a different combination of active and inactive sememe units.



IDEALIZED NEURON is the basis for artificial neural networks. It sums the weighted inputs that it receives from other neurons (*bottom*) and generates an activation level between 0 and 1. It then passes this activation (through weighted connections) to other neurons. The set of weights and connections in a neural network determines its behavior.

To make our neural network produce the correct pattern of semantic features for each word, we had to set the weight on each connection to the appropriate value. These weights are set not by hand but rather through a learning procedure—an algorithm for programming neural networks. To teach a network a task, one starts with random weights and then presents the network repeatedly with a "training set" of input patterns (in this case, letters in specified positions). The algorithm adjusts the weights after each training run to reduce the difference between the network's output and the "correct" response.

Neural-net workers have known since the 1950s how to adjust weights in simple, two-layer networks, but training networks with a greater number of layers is more difficult. In particular, it is not immediately obvious how to set the weights on the connections from the input units to the intermediate units because there is no way to determine, a priori, which intermediate units should be active for any given input and output.

During the 1980s, however, neural-net researchers developed a number of different methods for training multilayer networks. These methods apportion changes to the connection weights of each layer according to their contribution to the error. Over the course of many training cycles, the resulting weights converge to yield a network that produces the correct results. Depending on the initial random weights, learning may result in any of a number of sets of weights, each of which leads the network to produce correct answers for its training inputs. (For further details of the learning procedure, see "How Neural Networks Learn from Experience," by Geoffrey E. Hinton; SCIENTIFIC AMERICAN, September 1992.)

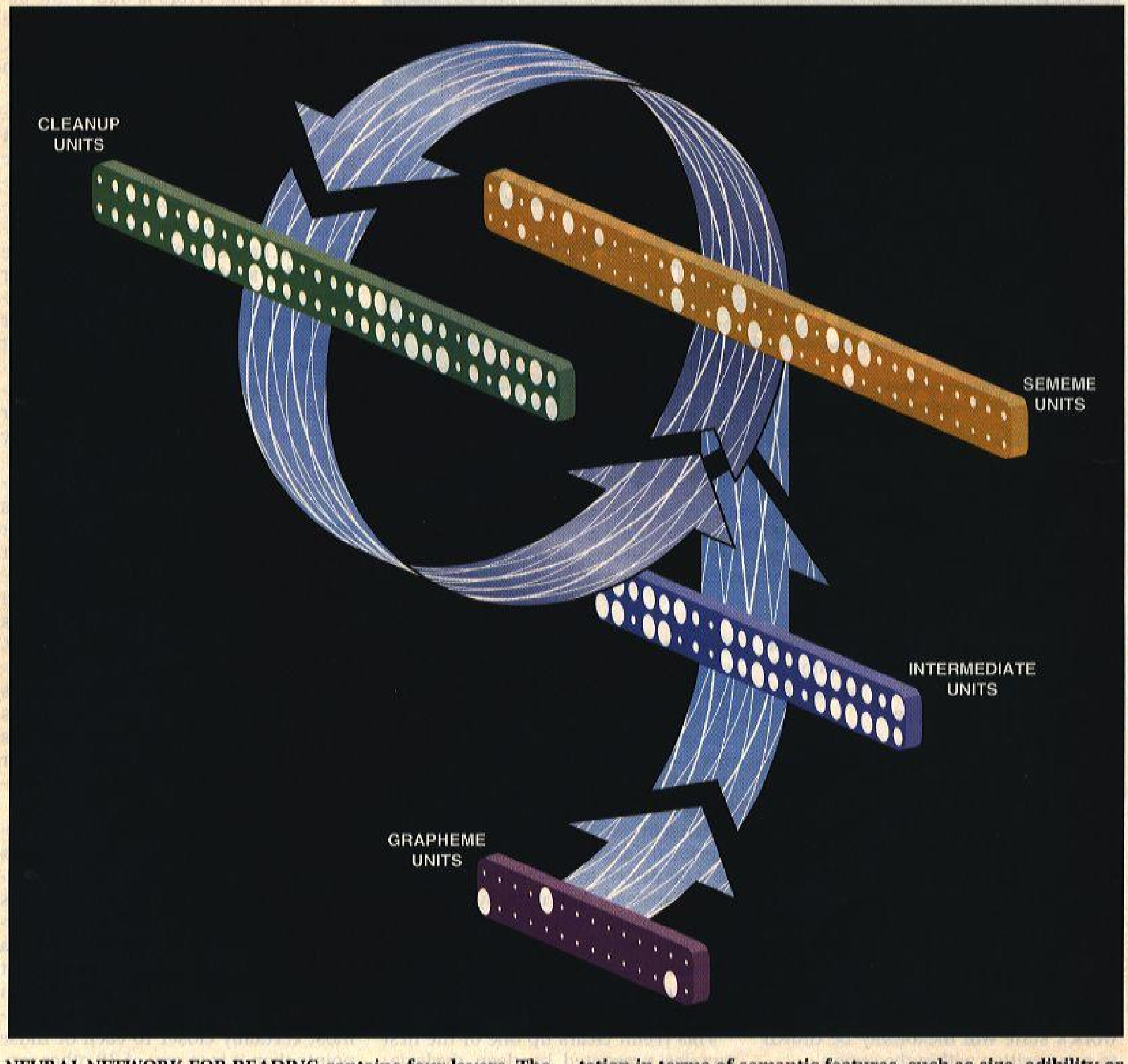
In theory, these learning procedures can get stuck in so-called local minima—configurations of weights that are incorrect but for which any small change would only make the network's errors worse. In practice, however, a network almost always learns nearly optimal solutions. In addition, some of the learning procedures are more biologically plausible than others, but our results do not seem to depend on which method we use. We suspect that even if the brain uses a quite different learning procedure, the resulting neural circuitry will still resemble the structure that our network develops. Thus, our explanation of what happens when the network is damaged may be correct even if its learning procedures are not.

Although our initial network, with one intermediate layer, could learn to map word-forms to their semantic features, it was not really satisfactory. It had a strong tendency to map very similar inputs (such as *cat* and *cot*) to similar outputs unless subjected to excessively long training. We addressed this problem by adding another layer of "cleanup" neurons. If the original set of connections produces a sloppy answer, the new units will change it to produce exactly the correct semantics. The number of word meanings is limited, so the pathway from the input need only get the activities of the sememe units closer to the correct meaning than to any other. The same learning techniques that succeed on networks with a single intermediate layer can direct the learning of nets containing multiple intermediate layers or even networks whose units are connected in cyclical fashion.

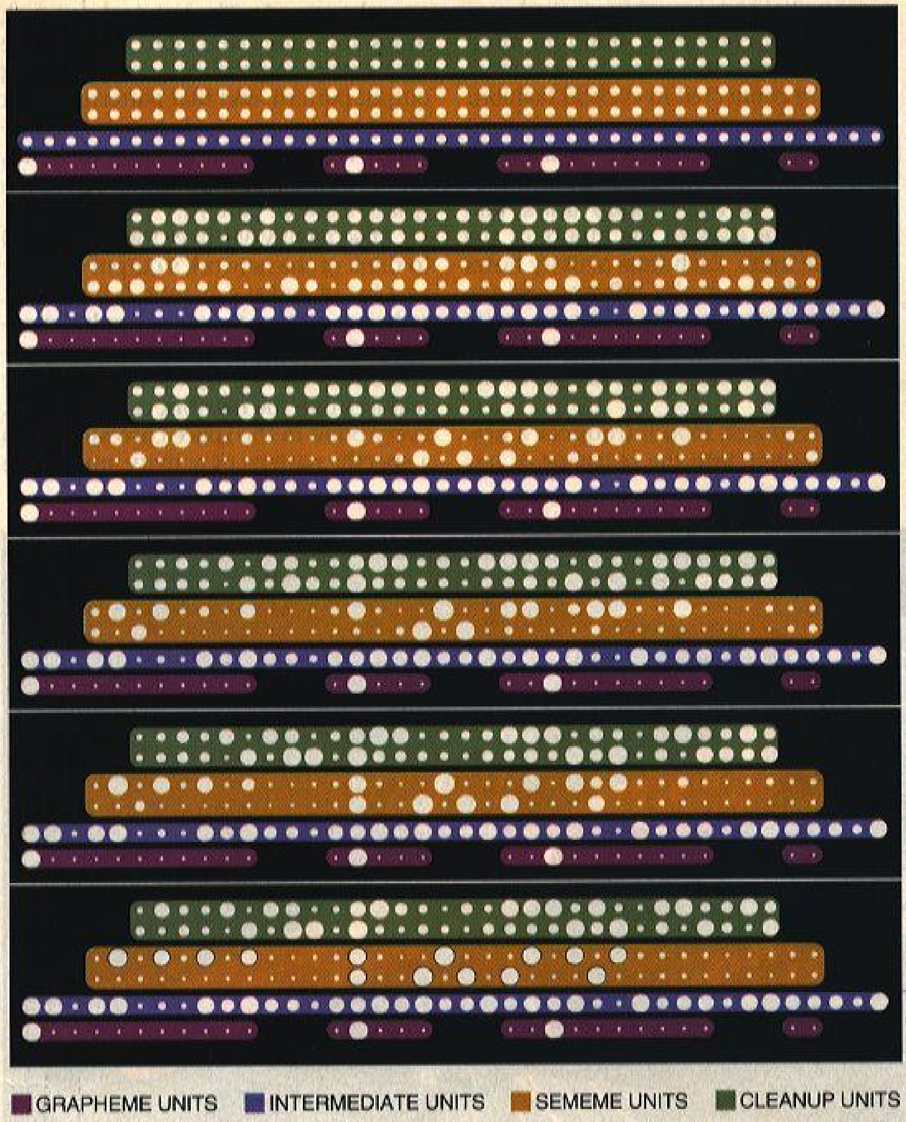
The most natural way to implement this cleanup mechanism is with a feedback loop. The output of the sememe units goes to the cleanup units, and their output goes to the inputs of the sememe units. Each time activity flows around the loop, the influence of the cleanup units on the sememe units (and vice versa) will yield a pattern of semantic features that is closer to the correct one.

The feedback loop introduces a new characteristic into the behavior of our neural network. The original network was static—any given input would cause the network to produce a corresponding output pattern, and that pattern did not change as long as the input stayed constant. The output of the new network, however, is dynamic; it settles gradually into a stable pattern.

Consequently, we have found it useful to think of the network's output not just as a list of active semantic features but rather as motion through a multidimensional "semantic space," whose coordinates are defined by all the semantic features that the network can represent. Every point in the space corresponds to a specific pattern of activity among the



NEURAL NETWORK FOR READING contains four layers. The first responds to the letters in each word. Connections between input and intermediate units and between intermediate and "sememe" units convert the word-form to a representation in terms of semantic features, such as size, edibility or aliveness. "Cleanup" units are connected to sememe units in a feedback loop that adjusts the sememe output to match the meanings of words precisely.



ACTIVATION LEVELS of neurons in the network change with time as the net processes the word *bed*. At first, many of the sememe units are activated to varying degrees, but interaction with the cleanup units strengthens the activation of some and weakens that of others until the output converges.

sememe units, but only a few of those patterns correspond to valid meanings. The correct meanings of words are points in semantic space.

The first three layers of the network, seen according to this perspective, take a word-form and convert it to a position somewhere in semantic space. Activity in the cleanup layer then draws the output of the network to the point corresponding to the closest meaning. The region around each word is what physicists and mathematicians know as a point attractor—whenever the network's initial output appears within a certain region, the network's state will inexorably be drawn to one position within the region.

This notion of a semantic space dotted with attractors representing the meanings of words has proved valuable for understanding how our network operates and how it can make the same semantic errors that dyslexics do. If we damage the network by randomly changing the weights in the cleanup mechanism, for example, the boundaries of the attractor for each word will change. As a result, if the network is in a region in semantic space where it was previously drawn to one word, it may now be drawn to a semantically related one instead. Alternatively, if we disrupt the pathway coming from the input, the network's initial output may be closer to the meaning of a semantically related word than to the meaning of the word originally presented.

This result clears up one of the first puzzles presented by deep dyslexia: why damage to any part of the brain's semantic route produces an essentially similar pattern of misreadings. Neurologists and others had wondered how damage near the input—the visual part of the reading system—could cause semantic errors. According to our models, these errors arise naturally as the cleanup neurons use semantic information to try to make sense of the output of the damaged earlier stages.

The notion of attractors helps to explain another anomaly in the data as well. Almost all patients who make semantic errors also make some visual errors—they confuse a word like *cat* with a visually similar word like *cot*. They do not, however, make the sounding-out errors of surface dyslexics ("loave" for *love* or "deef" for *deaf*). This invariable connection between semantic errors and visual errors is odd. Some patients must have damage solely to the later stages of their processing systems, and one would intuitively expect them to make only semantic errors.

After implementing our neural-network model, we discovered to our great surprise that damage to the semantic cleanup circuit sometimes caused visual errors. Retrospectively, we can understand why: the earlier layers of an undamaged network can afford to produce somewhat similar semantic outputs for the words *cat* and *cot* because the cleanup circuit will steer each to its proper meaning. But when the cleanup circuit is damaged and the shapes of each attractor change, the output of the sememe units may fall into the attractor for a visually similar but semantically unrelated word.

This explanation did not initially occur to us because it relies on the idea that the boundary of the attractor for *cat* can come very close to the one for *cot* even though the two words are semantically dissimilar. One would expect the attractors for many other meanings to come between those for *cat* and *cot*. In a two-dimensional space this intuition is correct: if we choose 40 points at random to represent word meanings and construct fairly compact attractors around each point, the attractors for dissimilar meanings will not come anywhere near one another.

It is very dangerous, however, to assume that the same is true in spaces that have many dimensions. Our network represents 68 semantic features in its sememe units, and so the attractors for each of its 40 words reside in a 68-dimensional space. It turns out that in 68 dimensions, the midpoint between any two randomly chosen points is almost certainly closer to each of those points than it is to any of 38 other random points. Consequently, the attractors for *cat* and *cot* can have a common border without any other attractors getting in the way. Avoiding obstacles is easy in 68-dimensional space.

Although our network was able to mimic both the correct and dysfunctional mapping of word-forms to meanings, that does not mean its architecture is the only possible one for the brain's semantic processing route. To determine the range of possible alternatives, we investigated the effects of damage on several different architectures, each designed to evaluate one aspect of the original network design.

We programmed versions of the neural network that contained connections among the sememe units and ones that lacked such connections; we also programmed some networks so that each neuron in one layer was connected to every neuron in the succeeding layer and others whose connections were sparse. In addition, we moved the cleanup units so that they performed their work ahead of the sememe units, and we combined the cleanup units with the intermediate layer. We even changed the arrangement of neurons in the input layer to alter the way that words were represented and added an output network that converted meanings to strings of phonemes, so that the system actually spoke.

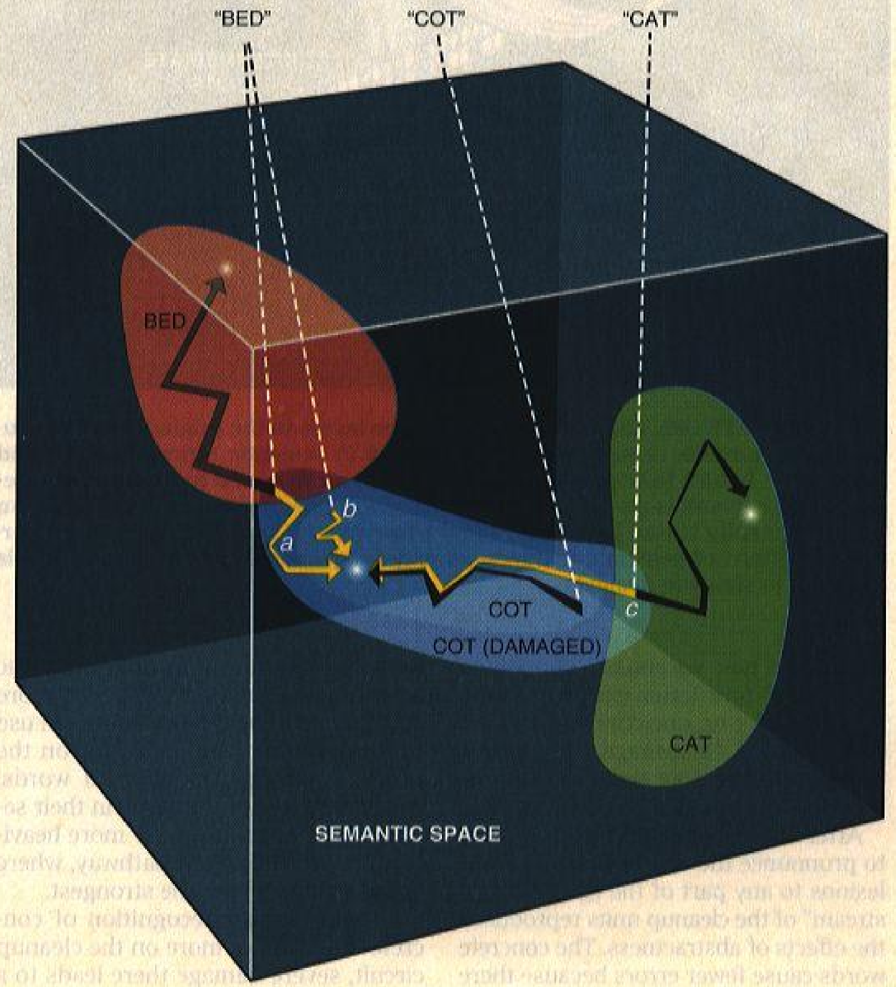
Most of the architectural details are irrelevant. The specific way the visual input is represented is not important as long as words that resemble each other visually produce similar patterns of activity in the input layer. The only crucial ingredient is the existence of attractors—if there are no cleanup units "downstream" of the damage, the network does not exhibit the pattern of errors characteristic of deep dyslexia.

Interestingly enough, our network not only reproduces the obvious visual and semantic errors of deep dyslexia, it also mimics some of the subtler characteristics of the disorder. For instance, patients occasionally make "visual then semantic" errors, in which a semantic confusion seems to follow a visual one. G.R. would read *sympathy* as "orchestra" (presumably via *symphony*). Our networks also produce these errors—sometimes reading *cat* as "bed," via *cot*.

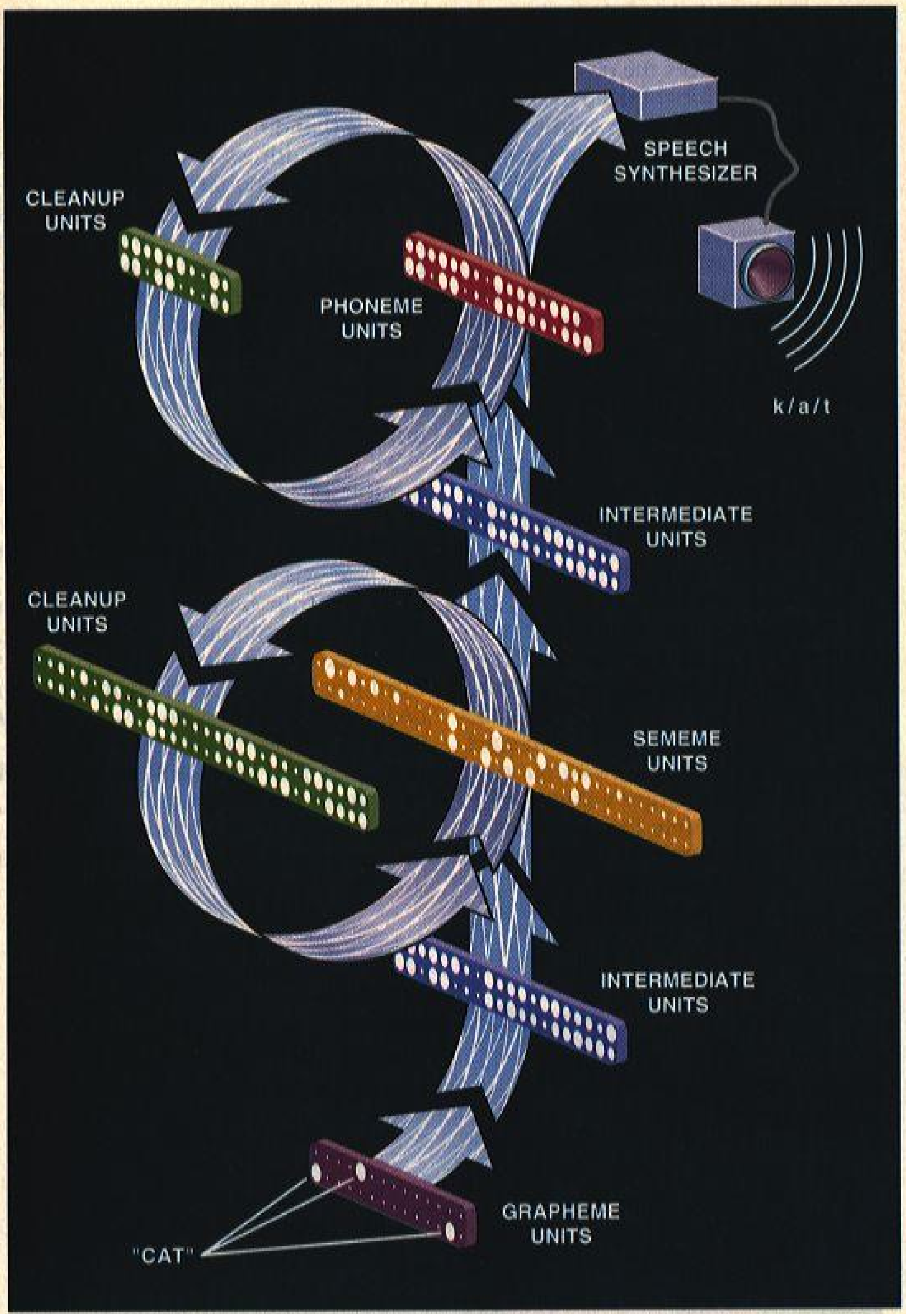
When severely damaged, our network also exhibits a strange effect that occurs when patients have a lesion so large that their semantic representations are distorted beyond recognition and they cannot find a word at all. Such patients are unable to identify the word they are trying to read, but they can often still decide which category it falls into, say, "animal" versus "food." Under similar circumstances, our network no longer stabilizes at the attractor corresponding to a particular word—indeed, the attractors for several words may have merged. Nevertheless, the network's output does stabilize within a larger volume of semantic space wherein the correct word and its relatives once resided. Consequently, the word's category can still be determined.

One symptom of deep dyslexia that our models did not initially address is the way in which patients have more trouble reading abstract words than concrete ones. This phenomenon appears to be an integral part of the syndrome because abstractness—a semantic property—increases the probability of visual errors. Furthermore, when patients make such misreadings, the responses they come up with tend to be more concrete than the original word presented.

We based our approach to simulating this effect on the proposal, made by Gregory V. Jones of the University of Warwick in England and others, that concrete words are easier for deep dyslexic patients because they evoke a more consistent and detailed meaning. In terms of our network, a concrete word has more semantic features than does an abstract one. For example, *post* has 16 features ranging from "size between one foot and two yards" to "used for games or recreation." In contrast, *past* has only two features: "has duration" and "refers to a previous time." We



SEMANTIC SPACE has many dimensions, corresponding to the semantic features (only a three-dimensional approximation is drawn here). The meanings of particular words are points in semantic space. When the authors' neural network reads a word, interaction between sememe and cleanup units causes any word-form that is mapped into a region of semantic space near the meaning of a word (*colored regions*) to converge on that meaning (*dots*). If the network is damaged so that the boundaries of these so-called attractors shift, a word can be misread as a semantically similar one—"cot" for *bed*, for example (*a*). Semantic errors may also occur if damage causes a word-form to be mapped to a slightly different point in semantic space (*b*). Such a network can make visual errors because visually similar words will initially be mapped to nearby points in semantic space, even if the stable points of the attractors they fall into are quite distant (*c*).



SPEAKING NETWORK adds another set of three layers to the original reading neural net. It converts sequences of letter shapes to semantic representations and maps those in turn to sequences of phonemes that can be fed to a speech synthesizer. This network is particularly useful because it does not require researchers to make potentially biased judgments about what word (if any) corresponds to a perturbed pattern of semantic features, as may be generated when the network is damaged to simulate dyslexia.

designed a new vocabulary containing 20 pairs of four-letter words differing by a single letter, one concrete and the other abstract. On average, the concrete words had about four times as many semantic features as did the abstract ones.

After the network had been trained to pronounce the words, we found that lesions to any part of the network "upstream" of the cleanup units reproduced the effects of abstractness. The concrete words cause fewer errors because there is more redundancy in their semantic activity patterns. Hence, there is more structure that the cleanup units can use to make the network converge on the proper meaning. The abstract words, which have less redundancy in their semantic patterns, must rely more heavily on the feed-forward pathway, where visual influences are the strongest.

Because correct recognition of concrete words relies more on the cleanup circuit, severe damage there leads to a surprising reversal: the damaged network reads concrete words less well and produces more visual errors than with abstract words. This type of lesion and pattern of performance are consistent with what is known about the single, enigmatic patient with "concrete-word dyslexia," studied by Elizabeth K. Warrington at National Hospital in London. Not only did he have much more trouble reading concrete words than abstract ones, he also did better matching spoken abstract words with pictures. This consistency suggests that his problem lay at the level of the semantic system.

Our account of the error pattern of deep dyslexia relies on the properties of a neural network that transforms one representation (a visual word-form) into another, arbitrarily related representation (a set of semantic features). One would expect similar error patterns to result from damage to other cognitive processes that involve an arbitrary transformation to or from a semantic space. Moreover, neuropsychologists have already described somewhat similar error patterns in deep dysgraphia, a disorder of writing, and deep dysphasia, a disorder of word repetition.

This additional evidence suggests that our model may have a wider validity than we originally supposed. More important, however, it marks the successful use of a new technique for understanding how the brain works. Our work differs from other explanations for deep dyslexia (and, with few exceptions, other explanations for neuropsychological phenomena in general) in the kinds of hypotheses that we frame. Instead of verbally characterizing each component in a complex neural mechanism and relying on intuition to tell us how damage will affect its behavior, we simulate that mechanism, damage it and watch to see what happens. We have found that many of our hunches were wrong. This discovery suggests that detailed computer simulations will play a crucial role in furthering understanding of how the brain normally processes information about language and of how that function is disrupted by injury or disease.

FURTHER READING

LESIONING AN ATTRACTOR NETWORK: INVESTIGATIONS OF ACQUIRED DYSLEXIA. G. E. Hinton and T. Shallice in *Psychological Review*, Vol. 98, No. 1, pages 74–95; January 1991.

DEEP DYSLEXIA: A CASE STUDY OF CONNECTIONIST NEUROPSYCHOLOGY. D. C. Plaut and T. Shallice in *Cognitive Neuropsychology*, Vol. 10, No. 5, October 1993.