COGNITIVE SCIENCE Vol 24 (1) 2000, pp. 151–167 ISSN 0364-0213

A Neuronal Basis for the Fan Effect

PHILIP GOETZ
Intelligent Automation, Inc.

DEBORAH WALTERS
SUNY at Buffalo

The *fan effect* says that "activation" spreading from a concept is divided among the concepts it spreads to. Because this activation is not a physical entity, but an abstraction of unknown lower-level processes, the spreading-activation model has predictive but not explanatory power. We provide one explanation of the fan effect by showing that distributed neuronal memory networks (specifically, Hopfield networks) reproduce four qualitative aspects of the fan effect: faster recognition of sentences containing lower-fan words, faster recognition of sentences when more cues are provided, faster acceptance of studied sentences than rejection of probes, and faster recognition of sentences studied more frequently. These are all a natural result of the dynamics of distributed associative memory.

I. INTRODUCTION

In many controlled experiments, the more facts a human knows about something, the longer it takes him or her to react to statements about that thing (Anderson, 1974, 1983a, 1983b). The ACT family of memory retrieval models (Anderson, 1983a, 1993) dubs this phenomenon the *fan effect*, and explains it via two suppositions:

1. Each object in memory ("cognitive units" in (Anderson, 1983a), "chunks" in (Anderson, 1993) has an activation associated with it, and the time taken to retrieve an object from memory is inversely proportional to its activation level.
2. Activation spreads from one object to another. The amount of activation an object X passes to each of the n objects it is associated with is inversely proportional to n. n is the *fan* of X.

In other words, the second supposition says that when spreading activation between concepts, a concept divides its activation among all the concepts connected to it.

Direct all correspondence to: Philip Goetz, 1A1, 2 Research Place, Suite 202, Rockville, MD 20850; E-Mail: flick@populus.net.

This conflicts with the usual model of a node or neuron in a neural network, such as Hopfield (Amari, 1972; Hopfield, 1982) or backpropagation (Rumelhart, Hinton, & Williams, 1986) networks, which compute a neuron's output as a nonlinear sigmoid or step function of its inputs, so that it tends to be either low or high. They do this to be able to compute nonlinear functions (Minsky & Papert, 1969). These models seem to me to be incompatible with the second supposition of the ACT model, and hence with the fan effect.

However, just because we're using the word "activation" for both the activity of simulated neurons, and for the activity of nodes in the ACT model, does not mean they are the same thing. Neurons do not exhibit the fan effect; the ACT model does. The ACT model represents a single concept with a single node; the brain probably does not. Anderson himself says, "At some level of abstraction it is reasonable to identify activation with rate of neural firing (although nodes probably do not correspond to simple neurons)" (Anderson, 1983a, p. 27). So ACT activation is not the same thing as single neuron activity. Rather, it must be an epiphenomenon of neuronal activity. This article shows that the fan effect is compatible with nonlinear sigmoid output functions by building a memory out of subsymbolic on/off neurons and showing that it exhibits the fan effect.

Suppose we have a set of lists of words to memorize, each list of length M. Assign a random N-bit binary string from $\{1, -1\}^N$ to each possible word. Represent a list of words $s = (w_1, w_2, \ldots w_M)$ as the concatenation of the N-bit strings for $w_1, w_2, \ldots w_M$. Use the Amari/Hopfield algorithm (Amari, 1982); same as Hopfield, 1982 modified to take activation values of 1 or -1 instead of 1 or 0) to store the lists in an MN-neuron Hopfield network as follows. r_{ij} = weight of connection (symmetric) between neurons i and j, V_i^s = activation (1 or -1) of neuron i in the pattern representing list s:

$$r_{ij} = \begin{cases} \sum_s V_i^s V_j^s & i \neq j \\ 0 & i = j \end{cases} \tag{1}$$

When the network is presented with a bit pattern (a vector in $\{1, -1\}^N$), each neuron readjusts its state repeatedly, setting:

$$V_i \leftarrow \begin{cases} 1 & \mathit{if} \sum_{j \neq i} r_{ij} V_j > \mathit{threshold} \text{ (0 in this paper)} \\ V_i & \mathit{if} \sum_{j \neq i} r_{ij} V_j = \mathit{threshold} \\ -1 & \mathit{if} \sum_{j \neq i} r_{ij} V_j < \mathit{threshold} \end{cases} \tag{2}$$

until no further neurons can change. Hopfield proposed that each neuron update itself randomly and asynchronously. In order to measure the time a Hopfield network takes to settle, we can let each node update itself once (in random order), and count that as one *relaxation iteration* or *cycle*.

Suppose you present the network with a corrupted version (obtained by randomly flipping the activity level of some nodes between 1 and -1) of the previously-stored target list ($w_1, w_2, \ldots w_M$). The fan of w_i is the number of other target patterns which have the word w_i in position i. Thus, the higher the fan of the words in a target pattern, the more competing patterns there are that overlap to a significant degree with the target pattern. Hence any dynamic system will take longer to stabilize into one pattern. Let's go back and look at some of the experiments used to suggest the fan effect, and see how this supposition lets us explain the data.

II. RECOGNITION AND RECALL

Experiment: The Basic Fan Effect

Anderson (1974) had 63 subjects memorize 26 facts of the form *A person is in a location*. Then they had to judge whether probe sentences were from the study set. The more study sentences that a person or a place were in, the longer subjects took to recognize sentences using that word.

We reran this experiment, replacing human memory with two 700-bit input buffers and one 700-neuron Hopfield network. One of the two input buffers stored a "person name", and the other stored a "location". (The "names" and "locations" were simply random 700-bit strings over $\{1, -1\}$). The Hopfield network represented a "convergence zone" (Damasio, 1990) where the two inputs would converge and be associated with each other via Hopfield learning. The purpose of using the convergence zone was to isolate the Hopfield learning network from the raw input, in order to test sentence recognition rather than word recognition. If we had simply used one Hopfield network made up of the two input buffers, a pattern consisting of one word from a pair and a random second word would already match half the bits in a memorized pattern. By contrast, the pattern produced in the convergence zone by presenting just one word from a pair would on average not match the memorized patterns associated with that word any better than a random pattern would. (This is explained further in Section 6.)

This network was not large enough to memorize the 26 patterns as specified in (Anderson, 1974).[1] Instead, we did a series of 1000 runs, in each run generating eight random persons and eight random locations, then generating 12 targets by randomly combining a person and a location, with the restriction that no target could have a location or person fan greater than three (for comparison with Anderson's data). The randomization was intended to avoid any peculiarities of a specific setup such as Anderson's.

On each run, the 12 targets were presented to the input buffers. Each bit P_i in the "person" buffer and L_i in the "location" buffer was connected to every neuron cz_j in the convergence zone by a random weight e_{ij}, half of which were 1 and the other half -1. (Since the person and location patterns were random and uncorrelated, we judged it acceptable to use the same weight matrix to transfer both into the convergence zone.) Each neuron in the convergence zone was set to:

TABLE 1
Comparison of Recognition Times for Humans and the Hopfield Network, Both Exhibiting the Fan Effect

Location fan	Recognition time (human/Hopfield net) Fan of person name		
	1	2	3
1	1.111 s/2.00 cycles	1.174 s/2.02 cycles	1.222 s/3.02 cycles
2	1.167 s/2.02 cycles	1.198 s/3.02 cycles	1.222 s/3.52 cycles
3	1.153 s/3.00 cycles	1.233 s/3.40 cycles	1.357 s/4.02 cycles

Notes. Linear correlation coefficient between time and sum of fans is .892 for humans, .987 for the Hopfield network. The true averages for the Hopfield data increase from left to right and from top to bottom, with a confidence level of $> (100 - 10^{-7})$ % for every such pairwise comparisons of averages. Time for humans is in seconds.

$$cz_j = \begin{cases} 1 & \sum_{i=1}^{700} P_i e_{ij} + L_i e_{ij} \geq 0 \\ -1 & \sum_{i=1}^{700} P_i e_{ij} + L_i e_{ij} < 0 \end{cases} \tag{3}$$

[This gave a very slight bias to 1’s over −1’s, which was judged acceptable since it would increase orthogonality between targets, and its only likely ill effect would be to decrease the Hopfield network’s efficiency slightly (Amit, 1989, p. 64)]. The convergence zone was trained on the resulting patterns using the Hopfield algorithm (Hopfield, 1982), with the exception that the self-weights r_{ii} were set arbitrarily to 87 (700/8) instead of to zero, as this was found to decrease errors. Presumably this is because most input bits were correct to begin with.

One degraded copy of each of the target patterns was then generated, with a 10% chance of flipping each bit. In this and all other experiments in this paper, degraded copies were presented to the network and recognition time was measured as the number of relaxation iterations until the network settled to a fixed state, and the answer was counted as correct if the stable pattern was closer (in Hamming distance) to the correct target than to any other target. The noise rate was set arbitrarily at 10%. (Other noise levels were used in the other experiments to demonstrate that no particular noise level was critical.) Table 1 compares the average time it took to settle in those cases for which the network got the right answer (72% or higher in all cases) to the time taken in Anderson’s experiment.

Experiment: Faster Response with More Complete Probes

Anderson (1983a, pp. 114–115) describes an experiment in which subjects learned to assign numbers to four-element (location-subject-verb-object) sentences such as *In the bank the lawyer cheated the doctor*, then were presented with probes consisting of all four elements, random subsets of three, or random subsets of two. All the elements came from

TABLE 2
Humans and Hopfield Networks both Recognize More Complete Probes Faster

Words in probe	Time (human)	Av. iters. to converge (network)	Fraction correct (network)
2	2.42s	2.82	.899
3	2.34s	2.24	.993
4	2.29s	2.04	1.000

The true averages for the Hopfield network are in the order given with a confidence level of $(100 - 10^{-1722})$ % (Z-values of 132 and 90 for the pairwise comparisons of adjacent averages).

one sentence. Any two words in a probe unambiguously specified a sentence, so that the complexity of pattern-matching might not increase with the number of elements in the probe. Subjects took less time the more sentence elements they were given.

Anderson interpreted this as meaning that the more concepts that are provided for recognition, the faster activation accumulates in the proposition node representing the sentence. We can also interpret this experiment as indicating that the more concepts that are provided, the closer we start in the space of neuron activations to the attracting pattern representing the sentence, and the faster we get there.

We repeated this experiment with a 100-neuron Hopfield network. The network memorized eight target patterns on each run, which was the most the network could store and still consistently recognize all targets. The target patterns were strings in $\{1, -1\}^{100}$. For each of the eight targets, 30, 000 probes were made by copying the target with a 10% chance of flipping each bit between 1 and -1, and then setting half, one quarter, or none of the bits to zero (inactive). This represented the cases where the subject was given two, three, and four of the words from the memorized target, respectively. This network did not use a hierarchical structure, since the probes and each part of the probes were all trained on equally. Table 2 gives the average of the time it took to settle in those cases for which the network got the right answer, and compares these results to those for Anderson's experiment.

III. EFFECTS OF LEARNING

Frequency of Study

Subjects recognize facts they have studied more frequently faster. The relevant factor in determining how quickly a sentence is recognized is not really the fan of each component of the sentence, but the relative frequency of study of the sentence with respect to other sentences containing the same elements (Anderson, 1976, 1983b, p. 27). We stated at the start of this article that ACT predicts that the activation spread from a node representing a word W to the node representing a sentence S containing W would be inversely proportional to the number of known sentences the word W was in. Actually, ACT* (Anderson, 1983a) and ACT-R (Anderson, 1993) both say that the activation is an increasing function of the number of times that sentence S was studied, and a decreasing function of the number of times that all the sentences that *W* was in were studied, though the details are different for each.

If we equate studying a sentence multiple times with entering it in the Hopfield study set multiple times, the Hopfield model predicts this also. Repeated study of one sentence will make the weights reflect that sentence's pattern more, at the expense of the other patterns stored in the network. If you trained one network once on every sentence and another network twice on every sentence, the two networks would behave identically, showing that the relevant factor is the relative and not the absolute frequency of study.

Experiment: Pairs with Different Frequencies. Anderson (1976, section 8.3, pp. 284–288) describes an experiment in which 24 subjects studied person-location sentences. Each sentence was either a "fan 2" sentence, meaning that each word in it appeared in one other sentence, or a "fan 4" sentence, meaning that each word in it appeared in three other sentences.

There were three types of fan 2 sentences. In one type, all three of the sentences with either of the two words from the sentence appeared three times. This is referred to as a probability 50% sentence, since given either the person or the location, one can guess the sentence correctly 50% of the time. In another, the sentence under consideration appeared four times, and the two sentences sharing a word with it each appeared two times. These were probability 67% sentences. In the third, the sentence appeared two times and the two sentences sharing a word with it each appeared four times. These were probability 33% sentences.

Of the fan 4 sentences, probability 50% sentences appeared three times, probability 33% sentences appeared two times, and probability 17% sentences appeared once. Table 3 gives an example of such a study set.

Because the fan effect is a function of relative probability of the identified proposition given its constituents, sentences which are matched in probability but differ in fan should have similar identification times.

We trained a 400-neuron Hopfield network on this task, making 24 random 200-bit patterns to represent the 24 distinct words used in a run of Anderson's experiment. We combined these patterns to generate 36 distinct target sentences, making one, two, three, or four copies of each pattern for a total of 72 targets for each network to memorize, as described in Table 3. Each word appears in six sentences, so the various cases are matched for word familiarity, and word recognition will not be confounded with sentence recognition.

Table 4 gives averages from 80 trial runs with 20% noise (20% of the bits flipped in each probe), and compares the results with Anderson's. Relative frequency of the words appearing in the probe sentence versus other sentences increases with "times studied." For both humans and Hopfield networks, recognition time is a decreasing function of relative frequency.

The important entries in this table are those in the probability 50% and 33% rows, for which we may compare sentences of different fan but identical probabilities given either of their constituents. As expected, recognition time does not in this case increase with fan.

TABLE 3
Structure of the 36 Subject-Verb Sentences

Example of material					
Fan 2 sentences			Fan 4 sentences		
		Probability (%)			Probability (%)
Baker	Street	50	Minister	Dungeon	50
Baker	Tavern	50	Minister	Orchard	17
Lawyer	Street	50	Minister	Barn	17
Lawyer	Church	50	Minister	Park	17
Farmer	Tavern	50	Hippie	Orchard	50
Farmer	Church	50	Hippie	Tower	17
Captain	Factory	67	Hippie	Dungeon	17
Captain	Swamp	33	Hippie	Barn	17
Banker	Boat	67	Artist	Barn	50
Banker	Factory	33	Artist	Dungeon	17
Fireman	Swamp	67	Artist	Hotel	17
Fireman	Boat	33	Artist	Orchard	17
			Sailor	Park	33
			Sailor	Tower	33
			Sailor	Orchard	17
			Sailor	Hotel	17
			Convict	Hotel	33
			Convict	Park	33
			Convict	Tower	17
			Convict	Barn	17
			Tailor	Hotel	33
			Tailor	Tower	33
			Tailor	Dungeon	17
			Tailor	Park	17

Note. From Anderson, 1976, Table 8.8, p. 285.

TABLE 4
Relative Frequency of Study Determines the Recognition Time of Sentences by Both Humans and Hopfield Networks

Recognition time (Human/Hopfield net) (Error rate) (Human/Hopfield net)			
		Fan of words in target	
Times studied	Probability	2	4
4	67%	1.369 s/2.065 cycles (.039/.000)	
3	50%	1.461 s/2.538 cycles (.053/.004)	1.520 s/2.201 cycles (.048/.000)
2	33%	1.542 s/4.888 cycles (.095/.721)	1.571 s/3.450 cycles (.070/.125)
1	17%		1.872 s/7.123 cycles (.262/.871)

Unfortunately, the opposite occurs. Recognition time with the Hopfield model also depends on the strength of competing memory traces. The fan 2 sentence “baker street” has one strong competitor for “baker,” the sentence “baker tavern,” which occurs equally as often. The fan 4 sentence “minister dungeon” has three competitors for the word “minister” but all three are weak competitors that appear only one-third as often as “minister dungeon.” These three competitors also compete against each other, and so the three combined make a weaker competition to “minister dungeon” than “baker tavern” does to “baker street,” and hence probability 50%/fan 4 sentences such as “minister dungeon” are recognized faster than probability 50%/fan 2 sentences such as “baker street.”[2]

It is possible that the discrepancy is related to the strategy used to accept or reject a probe (see Section 4). It is also possible that humans would exhibit a similar effect if trained on nonsense words.

Lag

The same human experiment examined the effects of lag on reaction time (Anderson, 1976, pp. 288–290). When you present a probe multiple times, reaction time increases as the number of other probes presented in between increases. If we modified our Hopfield model slightly, to suppose that whenever a probe is presented to the network it is also entered into the network (by adding it into Equation 1), and that weights decay over time, as in (Rumelhart & McClelland, 1986), then the more intervening items there were, the more the memory trace of the probe we are interested in would be crowded out by other traces, and the slower the network would be to converge on that pattern. We have not made this modification because it would make the other experiments in this paper more difficult, as we would then have to control for presentation order and introduce presentation of foils.

IV. REJECTION OF FOILS

In (King & Anderson, 1976), subjects memorized subject-verb-object propositions, and then judged whether the verb and the object in verb-object probes had appeared together in the same study proposition. Negative responses consistently took about 200ms longer than positive responses. A similar time difference was found in (Anderson, 1974).

The authors proposed that a subject waits for a set period of time for an intersection of activation spreading from the different words in the probe. If none occurs, the subject concludes that he has not seen the probe before. This model must make the waiting period longer when probes are constructed from higher-fan concepts, because higher-fan targets take longer to build up activation.

Suppose that when a subject is presented with a probe, the subject waits until something like a Hopfield network settles to a stable state. Define the *energy* of the resulting state as

$$-\sum_{i,j} V_i r_{ij} V_j \tag{4}$$

TABLE 5
Structure of the 16 Subject-Verb-Object Sentences

Schematic structure of sentences					
		Object Fan			
Verb fan		1		2	
1	Connected	$S_1V_1O_1$	$S_1V_2O_2$	$S_7V_5O_9$	$S_7V_6O_{10}$
	Unconnected	$S_2V_3O_3$	$S_3V_4O_4$	$S_2V_6O_{11}$	$S_3V_8O_{12}$
2	Connected	$S_4V_9O_5$	$S_4V_{10}O_6$	$S_{10}V_9O_9$	$S_{10}V_{11}O_{11}$
	Unconnected	$S_5V_9O_7$	$S_6V_{10}V_8$	$S_6V_{10}O_{10}$	$S_{12}V_{12}O_{12}$

Note. Derived from King & Anderson, 1976, Table 1.

This is the energy function that the Hopfield network is minimizing. Intuitively, we expect that the energy will be lower when the network converges onto or near to a target that it trained on than when it converges on a random pseudotarget. Spurious states are sometimes depicted as dimples on the surface of deeper, intentionally learned attractors (see, e.g., Amit, 1989, p. 82). But intuition misleads us. Empirically, the situation seems to be this: Training the network imposes some deep minima on it that do not correspond to any target. The targets are smaller, local minima that stud the search space. If you enter the space at a point that is not near one of these local minima, you fall into one of the deeper, accidental minima, taking longer and getting a lower energy score. So the situation is backwards from what we would expect: Recognized probes result in high-energy network configurations; foils result in low-energy configurations that take longer to find.

This suggests two alternate ways of rejecting a foil. One is, if the final energy is above a threshold, the probe has been recognized; if below threshold, reject as a foil. It does not require setting any special waiting period. The other is to set a waiting period, and reject the probe if the network does not settle within the waiting period.

We redid the experiment from King & Anderson, 1976 on a 270-node network, reducing the training set to 16 patterns because our computer did not have enough memory to learn 32 patterns well. The pattern of test "sentences" is given in Table 5.

The network memorized subject-verb-object triplets. (The presence of the subject position in King & Anderson, 1976 was to test a different property, called *connectedness*. The Hopfield experiments used probes constructed following the same connectedness pattern, to ensure comparable results.) The network was presented with 32 verb-object pairs, half of which occurred in the same study sentence, half of which were hybrid foils from two sentences. Probes were only 5% corrupted because the network was operating at about its maximum capacity. Both methods of foil rejection were tested. Hopfield Method 1 (200 runs of 32 probes each) rejected probes if the final energy level was less than $-1.23 \times \textit{neurons} \times \textit{neurons}$ (a value chosen empirically). Hopfield Method 2 (100 runs) rejected probes if the time taken to converge in cycles was greater than the verb fan plus the object fan. Table 6 compares the results with those of King & Anderson, 1976.

TABLE 6
Response Times and Error Rates by Fan and Type of Response (Positive or Negative)

	Object fan			
	"Studied"		"Not studied"	
Verb fan	1	2	1	2
1	.786/3.6/2 (.073/.34/0)	.853/3.9/2.67 (.067/.36/.28)	.998/9.6/3 (.161/.12/.18)	1.001/5.9/4 (.127/.42/.38)
2	.830/3.7/2.64 (.080/.33/.30)	.947/4.7/3.20 (.198/.62/.18)	1.014/5.5/4 (.124/.41/.38)	1.090/5.1/5 (.125/.51/.16)

Note. Human data from King & Anderson, 1976, Table 3. Time to give response, human (s)/Hopfield 1/Hopfield 2 (cycles). Error rate in parentheses, human/Hopfield 1/Hopfield 2.

The important thing to observe in this table is that all the entries in the "Not studied" column are greater than the corresponding entry in the "Studied" column. Table 7 shows the differences.

Using both methods, the Hopfield net did take longer to reject foils than to accept targets. A strange finding with Method 1, contradicting the data in Anderson (1974) and King & Anderson (1976), was that the fan effect worked backwards for rejected foils; foils of high fan were rejected faster than foils of low fan. This may be related to the tremendous error rates (near random) for high-fan targets and foils using Method 1.

Thus, when using the rejection strategy proposed in (King & Anderson, 1976), namely, rejecting a probe after a set waiting period dependent on the fan of the words involved, the Hopfield data matches the human data much better than it does when using the strategy of waiting until the network settles and examining its energy level.[3] So, the rejection strategy proposed in (King & Anderson, 1976) to match the human data is also an obvious strategy to use in conjunction with Hopfield networks, if we interpret "an intersection occurs" as "the network converges", and matches the human data qualitatively.

V. OBJECTIONS TO THE MODEL

Brains are not Hopfield networks. Why should we expect results using Hopfield networks to tell us anything about the brain?

The Hopfield network is one of a more general class of networks called *attractor neural networks* (Amit, 1989), which are in turn a subset of the class of *associative networks* (or

TABLE 7
Ratio of Time Taken to Respond "Not studied" versus to Respond "Studied"

	Object fan	
Object fan: Verb fan	1	2
1	1.27/2.7/1.5	1.17/1.5/1.5
2	1.22/1.5/1.5	1.15/1.1/1.6

Note. Human data from King & Anderson, 1976, Table 3.

auto-associative networks). Associative neural networks, not to be confused with associative networks of the symbolic type (Findler, 1979), are content-addressable memories that can recover a pattern from a part of the pattern, and are often composed of sets of neurons connected recurrently and homogeneously. Attractor networks are associative networks designed so that the target patterns are attractors of the network. Interference with other patterns occurs when attractors are close to each other. Given a probe vector near two target attractors, the competing attractors will both pull on the probe vector, so it will approach its final destination more slowly than if only one attractor were nearby. So we expect to find the same effects in any attractor memory.

Sparse, Asymmetric Connections

The basic Hopfield network is fully-connected, meaning that every node is connected to every other node, and symmetric, meaning that $r_{ji} = r_{ij}$ (see Equation 1). Brains are probably neither (Minai & Levy, 1993). Furthermore, symmetrically-connected, serial Hopfield networks converge to a single fixed pattern (Bruck, 1990). This makes them unsuitable for generating temporal patterns for sequencing behavior (Lukashin et al., 1996), especially chaotic dynamics (Skarda & Freeman, 1987; Celletti & Villa, 1996).

The Hopfield network has been altered to recall temporal sequences, by using asymmetric weights (Amari, 1972) and by using chaotic neurons (Adachi & Aihara, 1995). Also, the fan effect has been demonstrated only in tasks relying on declarative memory, for which convergence may be appropriate. Hubert (1993) showed that a 36-node Hopfield-like network with each node connected only to its eight neighbors was able to recognize its training set, though not as well as the fully-connected network. Goetz's unpublished experiments suggest that sparsely connected Hopfield networks are more efficient memories than are fully-connected ones, and that asymmetric networks are only slightly less efficient than symmetric ones.

Time of Presentation

Two effects of presentation order on free recall are the primacy effect (improved recall of words presented at the beginning of a list) and the recency effect (improved recall of words presented at the end of a list) (Bjork & Whitten, 1974). The primacy effect could be modelled by assuming that there are limits to the plasticity of a network's connections. If connections approach strong positive or negative connection strengths asymptotically, then the first patterns presented would have a greater effect on the memory trace than later patterns. The recency effect could be modelled by having connection strengths decay back towards their original values, which should favor the items most recently learned.

Parsimony

The word "lawyer" ought to activate the same units in representing "The doctor knew the lawyer" that it does in "The lawyer knew the doctor." Otherwise, the system would lack productivity of thought, systematicity of cognitive representation, and compositionality of representations (Fodor & Pylyshyn 1988). This is possible; Hinton has demonstrated a network to gate patterns into different parts of a larger attractor network (Hinton 1981, section 10).

Hierarchical Structure

The brain represents sensory information at many levels, passing it through a series of brain areas along multiple diverging, converging, and re-entrant pathways (Sporns et al., 1994; Kosslyn & Koenig, 1992, p. 54–60; Martin et al., 1995). There are also cognitive reasons for believing the brain has hierarchical structure, such as that, contrary to the fan effect, people respond faster to questions about things they are expert in (Hayes–Roth, 1977), and that the fan effect diminishes when learned facts are consistent with each other (Smith et al., 1978; Reder & Anderson, 1980).

The literature provides numerous ways to combine associative networks (Minsky, 1986; Damasio, 1990) and attractor networks in specific (Geman, 1981; Amit, 1989, chap. 8; Cartling, 1996) into hierarchies. The method used in the first experiment is reminiscient of (Geman, 1981) and (Damasio, 1990).

Memory Capacity

It has been claimed that the Hopfield network has serious memory capacity limitations (Wasserman, 1993; Kung, 1993; Forrest & Wallace, 1995). Each pattern in an n-node network takes $O(n)$ bits to specify. The network contains $O(n^2)$ bits; hence, the best capacity possible (in terms of the number of patterns that can be stored and recalled) is $O(n)$. The number of patterns which can be stored perfectly and recalled in one iteration is only $\frac{n}{2ln(n)}$ (McEliece et al., 1987; Gardner, 1987; Amit, 1989). But the proof computes the maximum number of patterns such that we have a fixed probability of all N bits of a stored pattern being stable. Small networks have an advantage because the minimum amount a pattern can be wrong by is one bit. If we allow a number of errors that is proportional to the network size, or allow a longer convergence time, the capacity is then $O(n)$ (McEliece et al., 1987; Gardner, 1987).

Spurious Responses

The Hopfield network sometimes produces spurious responses. In particular, the inverse of every target also becomes an attractor. But we do not know whether this is a bug or a feature. Although people often have considerable difficulty reading a page that has been turned upside down, they have no difficulty reading white letters on black computer monitors, even if they have read only black letters on white paper before.

VI. TOWARDS A TESTABLE HYPOTHESIS: THE FAN RATIO EFFECT

Many variations of the Hopfield experiments performed here have shown that fan ratio (the ratio between the fan of the various components of a target to be recognized) can have at least as large an effect on error rates and, under some circumstances, recognition times, as does total fan, causing large errors and slow recognition when the disparity in fan between components is large.

We originally ran the experiment of Section 2 using a single Hopfield network to store both parts of the pattern. This experiment has the highest fan ratio (three to one) of any of the experiments. The error rate and recognition time for a sentence whose components had fans of three and one were higher than for fans of three and two, or three and three, contrary to the predictions of the fan effect. The fan ratio effect had washed out the fan effect. The convergence zone was then used in that experiment to separate sentence recognition from word recognition; it also reduced the fan ratio effect.

Consider a single neuron, a_i, in the left half L of a two-part pattern, each part of which consists of n one-bit neurons. Suppose the right half of the pattern, R, has a fan of f_R, meaning it is associated with f_R different left halves. Suppose the network learned p left-right patterns in all.

Each pattern learned casts one bit-vote on whether the correlation between a_i, in the left half, and a_j, a neuron in the right half, is positive or negative. Suppose the pattern R is present in the right half of the network. Then R casts np votes on a_i's next state, 1 or -1. For each of the f_R patterns associated with R, R casts n identical votes. The other votes are random, uniformly distributed with a mean of zero, assuming that the p learned patterns are random with respect to each other and have an equal number of 1 and -1 bits.

Thus we see that putting R in the right half of the pattern influences the left half towards the f_R different patterns associated with R. The random (but constant) "background noise" from the $p - f_R$ other patterns, and the random correlations between the f_R possible left halves, will decide which pattern in the left half is preferred by R. If the left half were initialized randomly, it would have a one in f_R chance of settling on the correct pattern.

Now consider the effect of f_L, the fan of the left pattern, in the situation where L and R are both present, but corrupted by noise. If $f_L = 1$, even a degraded L will immediately cause the right half to converge on R. R will return the favor by causing the left half, if it is not yet firmly settled, to converge on its favorite of its f_R associates.

If $f_L > 1$, the right half of the network will receive conflicting information about which pattern it should converge onto. The greater f_L is, the longer the right half will take to converge. Error is least when $f_L = f_R$, because then both halves take roughly the same time to converge, and neither has the opportunity to transform the other into its favorite partner.

This explains why the use of the convergence zone reduces the fan ratio effect. The re-use of a person pattern P no longer causes half the training pattern to be identical in each sentence containing P. Rather, it causes the sentences containing the person P to bear some family resemblance to each other.

TABLE 8
Comparison of Error Rates for Humans and the Hopfield Network

Error rate (human/Hopfield net)			
	Fan of second proposition		
Fan of first proposition	1	2	3
1	.073/.000	.067/.004	.064/.292
2	.080/.001	.198/.013	.177/.040
3	.107/.269	.242/.052	.220/.012

Notes. Human errors are taken from King & Anderson, 1976, Table 3.

The degree to which fan ratio affects the error rate increases with the noise level in the probes, and with the strength of a pattern component's preferences for one association over others. It might be reduced or eliminated by perceptron learning (Amit, 1989, section 9.2) that trains all patterns until they are recognizable, or by unlearning (Hopfield, 1983) to accomplish the same thing.

The data in (King & Anderson, 1976, Table 3), from an experiment similar to that in (Anderson, 1974), gives a correlation coefficient between error rate and the sum of verb and object fan of .80 for positive responses, and .60 for negative responses. (The same correlation in the two experiments in (Anderson, 1974) was insignificant. This may be attributed to the difference in procedure: in (King & Anderson, 1976) the subjects studied each sentence once; in (Anderson, 1974), propositions that the subject made errors on during the study phase were studied until the subject got them right, thus equalizing error rates across all conditions.) Table 8 compares the error rate for positive responses reported in (King & Anderson, 1976, Table 3) to that of the Hopfield network in the first experiment, above. Note that the greatest error rates for humans are in the lower right-hand corner, while the greatest error rates for the Hopfield networks are in the lower left-hand and upper right-hand corners (where fan ratio is greatest). This shows that the fan ratio effect overpowers the fan effect at a fan of three, making the error rate data disagree with the human data.

The use of a convergence zone has not banished the fan ratio effect, merely pushed it beyond the range tested in humans. With the convergence zone and the parameters used in the basic fan effect experiment (Section 2), fan ratio effects appear at fans of four to one and five to one. There are several ways to reconcile the hypothesis that the fan effect is explained by neuronal mechanisms similar to Hopfield networks with the appearance of the fan ratio effect in Hopfield models:

1. Show that the fan ratio effect appears in humans at higher fans. Care must be taken both to provide noise in the recognition input, such as by presenting degraded probes, and to present the probe pattern briefly, as in these models, so that the continued presence of the probe does not prevent the pattern in memory from shifting away from the pattern presented.
2. Show that use of a larger network can banish the fan ratio effect. If, as hypothesized, the fan ratio effect is aggravated by coincidental covariance in "background noise"

that causes one pattern to be preferentially linked with another, then the effect should be reduced by reducing the variance in the correlations between patterns. This could be accomplished by using larger networks (unfortunately beyond our computational capacity at this time). The converse is true: Running the first experiment with only 100 instead of 700 neurons in each half of the pattern and in the convergence zone leads to a pronounced fan ratio effect. The largest network used in this work contained 700 neurons; for comparison, visual area V1 contains between 140 and 234 million neurons (Orban, 1984).

3. Show that the data being recognized in human experiments is sufficiently remote from the input that the fan ratio effect disappears, as it does for fans of up to 3 in the first experiment when a convergence zone one level removed from the input level is used.

VII. SUMMARY

The Fan Effect is Compatible with Associative Distributed Memory Networks

The Hopfield model produces the same qualitative timing results as humans do for the retrieval tasks examined. Doubtless human memory is more complex. But these results suggest that an associative distributed network which recognizes patterns by relaxation may be at the root of human memory.

Perhaps any distributed gradient search method, such as spin glasses or the model in (Rumelhart & McClelland, 1986), would produce similar results. A distributed gradient search converges slower when the initial point is near several local maxima, because the different local maxima compete, and the network may be pulled in several directions at once by different maxima winning along different dimensions of the search, until the distributed elements finally “agree” on one maxima to pursue. But gradient search requires a complex search space that symbolic representations do not provide. Breaking the symbolic representation down into a distributed representation provides the key to understanding the dynamics of recognition.

The Fan Effect is an Epiphenomenon

Rather than conflicting with nonlinear activation transfer functions, the fan effect may be an artifact of them. This doesn’t mean that the fan effect isn’t real. It is a simple way of describing a phenomenon that has complex neuronal causes. However, the fan effect is an abstract description, and the “activation” it talks about is also an abstraction. This paper shows that the fan effect may be caused by the spread of activation at a level of abstraction below that of the concepts under investigation in Anderson’s models. The fan effect is not incompatible with nonlinear search methods that switch quickly between different stable local maxima, as it at first appears.

Symbolic and Connectionist Models: Two Paradigms are Better Than One

The symbolic models of John Anderson and others described the fan effect and suggested experiments that further elucidated it. But without a neuronal explanation, the models remain enigmatic, describing but not explaining. In addition to providing a causal explanation, neuronal explanations may also restrict model development. Using both approaches is the best path to satisfactory models of cognition.

VIII. ADDENDUM

While writing this article, we found that Rumelhart and McClelland (1986) reported that Kevin Singley demonstrated the same basic idea in 1985 while he was a graduate student under John Anderson. Apparently he left academia before publishing the results (McClelland, personal communication 1996).

NOTES

1. Incorrect answers usually resulted form settling to a pattern that shared a person or location name with the desired target. That the patterns associated with some sentences were strongly correlated with each other explains why the network could not store .138 N patterns ($N = 700$), as expected when the memorized targets are randomly correlated (Amit et al., 1985).
2. There is also a problem with the limited "learning resolution" of the network. Any network large enough to recognize most of the sentences studied few times, was able to memorize the other sentences perfectly and recognize them in the minimum possible time of two cycles.
3. Both methods present difficulties for neural implementation: it is not clear how to calculate either the energy level of a network or the fan of a word using neurons.

REFERENCES

Adachi, M., & Aihara, K. (1995). Associative dynamics in a chaotic neural network. Mathematical Engineering Technical Report 95-01, University of Tokyo.

Amari, S. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers, C-21,* 1197–1206.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters 55*, 1530–1533.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge, UK: Cambridge University Press.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology 6*, 451–474.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1983a). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1983b). Retrieval of information from long-term memory. *Science, 220,* 25–30.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*, 173–189.

Bruck, J. (1990). On the convergence properties of the Hopfield model. *Proceedings of the IEEE, 78,* 1579–1585.

Cartling, B. (1996). Dynamic control of semantic processes in a hierarchical associative memory. *Biological Cybernetics, 74*, 63–71.

Celletti, A., & Villa, A. E. P. (1996). Low-dimensional chaotic attractors in the rat brain. *Biological Cybernetics, 74*, 387–393.

Damasio, A. R. (1990). Synchronous activation in multiple cortical regions: A mechanism for recall. *The Neurosciences, 2*, 287–296.

Findler, N. (1979). *Associative networks: Representation and use of knowledge by computers*. New York: Academic Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture. *Cognition, 28*: 3–71.

Forrest, B. M., & Wallace, D. J. (1995). Storage capacity and learning in Ising-spin neural networks. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks* (Vol 1, pp. 129–156), Berlin: Springer–Verlag.

Gardner, E. (1987). Multiconnected neural network models. *Journal of Physics A, 20*, 3453–3464.

Geman, S. (1981). Notes on a self-organizing machine. In G. E. Hinton & J. A. Anderson, (Eds.), *Parallel models of associative memory* (pp. 275–303). Hillsdale, NJ: Lawrence Erlbaum.

Hayes–Roth, B. (1977). Evolution of cognitive structures and processes. *Psychological Review, 84*(3), 260–278.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 187–217), Hillsdale, NJ: Lawrence Erlbaum.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences 79*, 2554–2558. Reprinted in J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: Foundations of research*, Cambridge, MA: MIT Press.

Hopfield, J. J. (1983). 'Unlearning' has a stabilizing effect in collective memories. *Nature, 304*, 158–159.

Hubert, C. (1993). Design of fully and partially connected random neural networks for pattern completion. *New trends in neural computation: International workshop on artificial neural networks '93*. Berlin: Springer-Verlag.

King, D. R. W., & Anderson, J. R. (1976). Long-term memory search: An intersecting activation process. *Journal of Verbal Learning and Verbal Behavior, 15*, 587–605.

Kosslyn, S. M., & Koenig, O. (1992). *Wet mind*. New York: Macmillan.

Kung, S. Y. (1993). *Digital neural networks*. Englewood Cliffs, N.J.: Prentice Hall.

Lukashin, A. V., Amirikian, B. R., Mozhaev, V. L., Wilcox, G. L., & Georgopoulos, A. P. (1996). Modeling motor cortical operations by an attractor network of stochastic neurons. *Biological Cybernetics, 74*, 255–261.

McEliece, R. J., Posner, E. C., Rodemich, E. R., & Venkatesh, S. S. (1987). The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory, IT-33*, 461–482.

Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science 270*, 102–105.

Minai, A. M., & Levy, W. B. (1993). The dynamics of sparse random networks. *Biological Cybernetics, 70*, 177–187.

Minsky, M. (1986). *The society of mind*. New York: Simon and Schuster.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Orban, G. A. (1984). *Neuronal operations in the visual cortex*. Berlin: Springer-Verlag.

Reder, L. M., & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. *Cognitive Psychology, 12*, 447–472.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 2). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). A distributed model of memory. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 2). Cambridge, MA: MIT Press.

Skarda, C. A., & Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences, 10*, 161–195.

Sporns, O., Tononi, G., & Edelman, G. (1994). Reentry and dynamical interactions of cortical networks. In E. Domany, J. L. van Hemmem, & K. Schutter (Eds.), *Models of neural networks* (Vol. 2, pp. 315–341). Berlin: Springer-Verlag.

Wasserman, P. D. (1993). *Advanced methods in neural computing*. New York: Van Nostrand Reinhold.