Representing causal schemata in connectionist systems

Richard M. Golden

Department of Psychology

Brown University

The connectionist approach to human memory is based upon the idea that knowledge can be stored implicitly in the form of real-valued interconnections among a set of simple "neuron-like" computing elements (Hinton & Anderson, 1981). The schema system approach (Rumelhart, 1980; Schank & Abelson, 1977) considers human memory to be organized in terms of many small packets of knowledge called schemata. If a knowledge packet is defined as some sequence of causally related events, then it is referred to as a "causal schema" or "script."

Although these two seemingly different approaches to the problem of modelling human memory might seem incompatible, they are actually intimately related (Rumelhart, Smolensky, McClelland, & Hinton, 1986; also see Touretzky & Hinton, 1985). In this paper, a connectionist model of how causal schemata are used in the recall of actions from simple stories is described. The paper is organized in the following manner. First, an explicit procedure for representing complex causal schemata as "neural activation patterns" is discussed in detail. Next, the fundamental neural mechanisms that are used to process and learn information are described and motivated from a probabilistic viewpoint. Finally, the resulting system is used to model some experimental data obtained by Bower, Black, and Turner (1979) in their studies of human memory for written text.

Representational assumptions

The fundamental entity in this model is called a "causal relationship." A Causal Relationship (CR) consists of an "initial situation," an "action," and a "final situation." If the "final situation" of one causal relationship is identical to the "initial situation" of another causal relationship, then the pair of causal relationships are said to be "causally linked." A collection of causal relationships that have been linked together in this manner is referred to as a causal schema. For example, let the notation (S3,A3,S5) indicate a causal relationship formed from an initial situation S3, an action A3, and a final situation S5. Figure 1 shows that CR (S1,A1,S3) is causally linked to CR (S3,A3,S5) which, in turn, is linked to CR (S5,A2,S6).

A basic assumption of the model discussed here is that a causal relationship corresponds to some unique pattern of neural activity in the brain. In addition, similar causal relationships are assumed to possess similar neural codings. More specifically, a causal relationship is represented as a 160-dimensional state vector (i.e., a list of 160 real numbers) where the ith element of the vector specifies the firing rate of the ith neuron in a neural network. Consider the causal relationship at the top of Figure 2. The initial situation field of this CR is interpreted as: "The actor is at a restaurant, the actor is hungry, and the actor is at the table." The action field of the CR is interpreted as: "The actor orders the meal." The final situation field is interpreted as: "The actor is at the table, the actor is hungry, the food is on the table."

To encode the initial situation field as a 64-dimensional subvector, the three 64-dimensional binary orthogonal subvectors corresponding to the states: "At_restaurant," "At_table," and "Hungry" are added together to form a composite 64-dimensional subvector. If an element of this composite subvector is non-negative, then the value of that element is set equal to +1, otherwise the value of that element is set equal to -1. The resulting modified composite subvector represents the initial situation field of the 160-dimensional causal relationship state vector. More formally, the 41 states in the "state dictionary" form a psychological *basis set* that

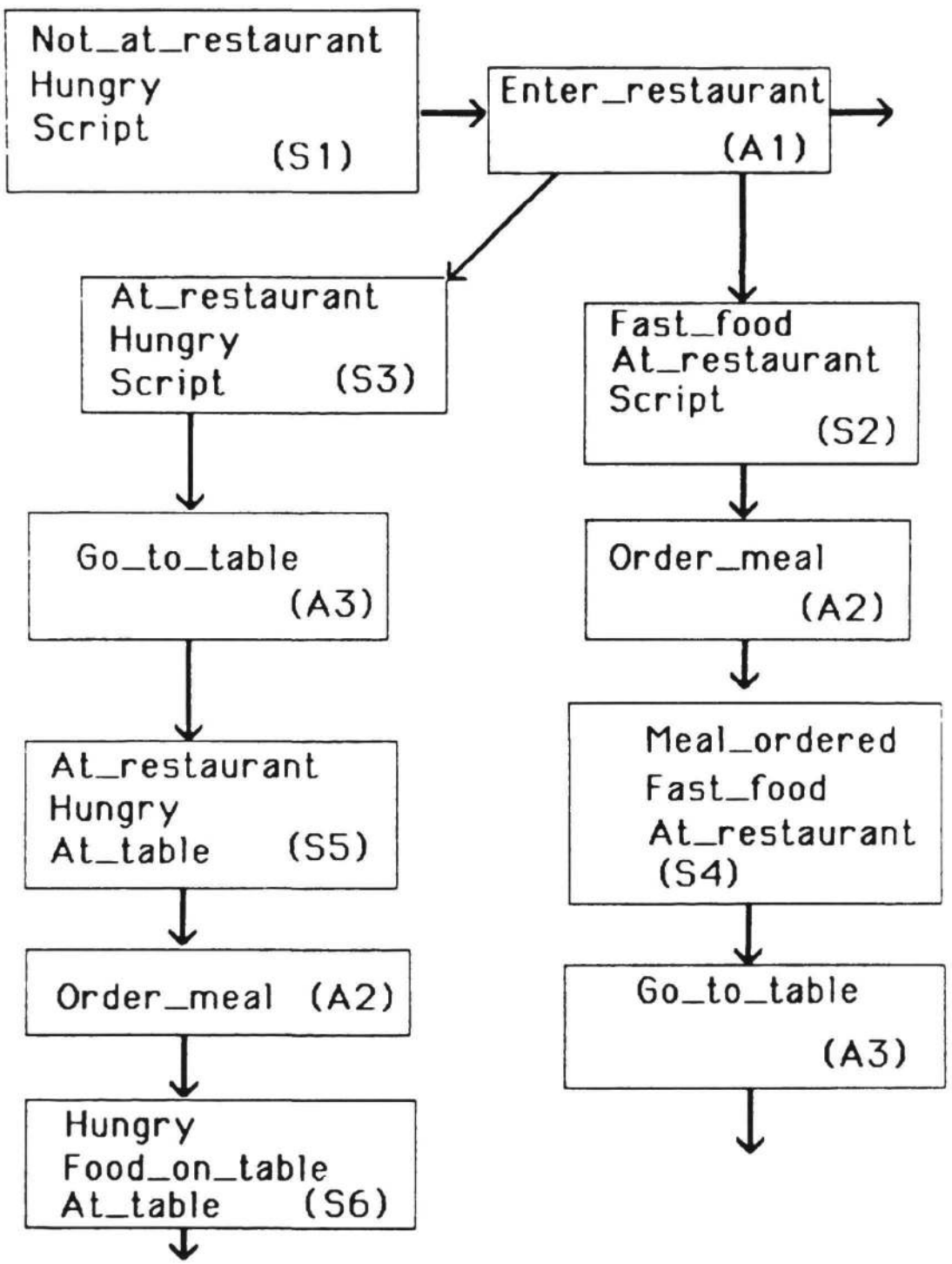

Figure 1. A portion of a causal schema. Note that such schemata may be represented as unordered collections of causal relationships.

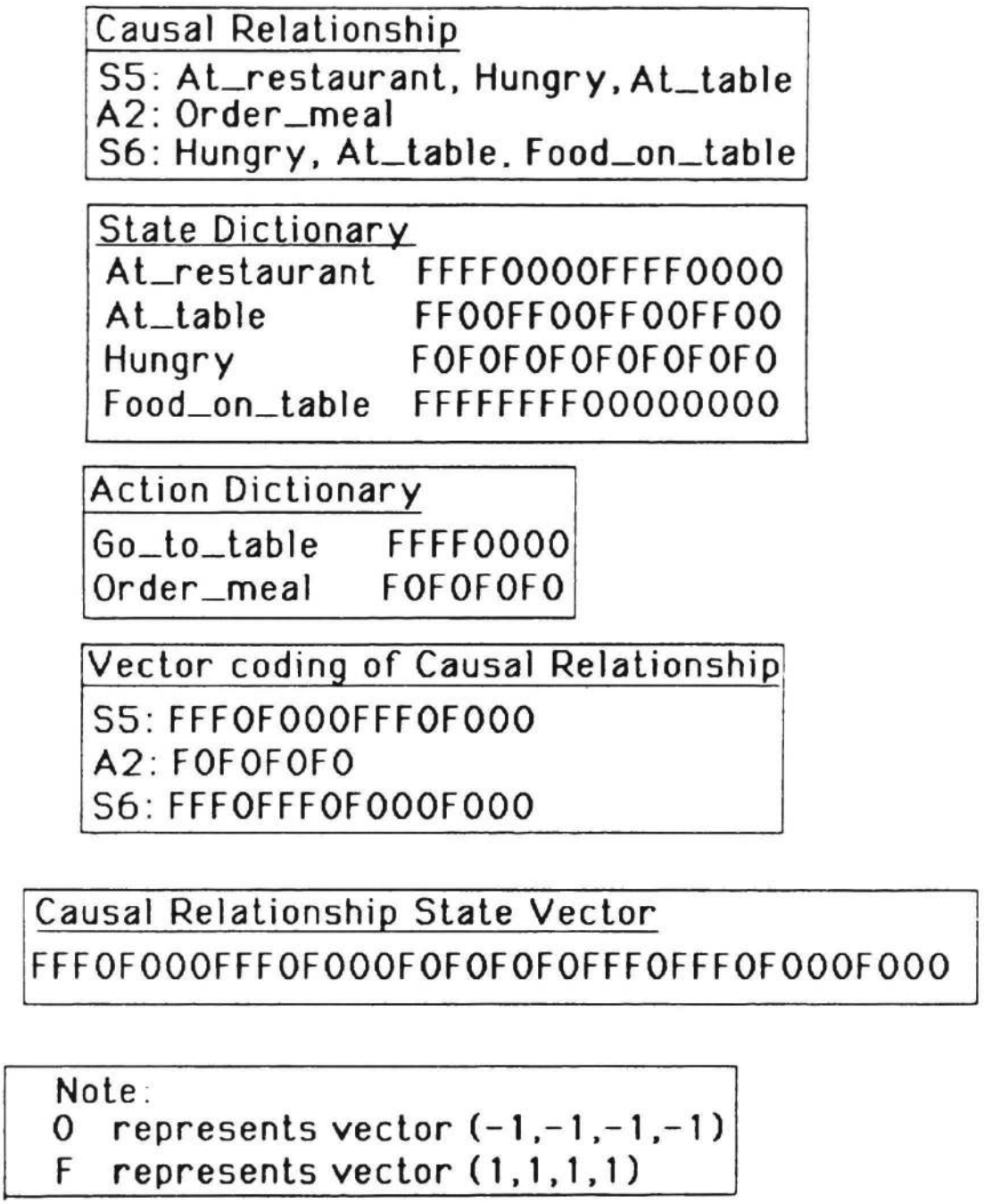

Figure 2. Representing a causal relationship as a state vector. The initial situation field of the CR is formed by adding the 64-dimensional subvectors in the state dictionary labelled "At_restaurant," "Hungry," and "At_table" together and assigning a +1 to the non-negative elements of the resulting vector and a -1 to the negative vector elements. The action field of the CR is simply looked up in the action dictionary. The final situation field of the CR is constructed in the same manner as the initial situation field. Note that the symbol F refers to a sequence of four positive ones, while the symbol 0 refers to a sequence of four negative ones.

can represent over 50,000 situations in a 64-dimensional state vector space.

The action field of the causal relationship is represented by a 32-dimensional binary orthogonal subvector whose value is obtained directly from the "action dictionary." The encoding procedure for the final situation field of the causal relationship vector is identical to the procedure used to encode the initial situation vector field.

Making "most probable" decisions with a neural model

The fundamental problem of content-addressable memory may be formulated as follows. Given some unusual or "improbable" vector $\mathbf{X}_0$, construct a more probable interpretation. More formally, we can search for a maximum of some probability density function P(**X**) in the vicinity of $\mathbf{X}_0$. This density function indicates the relative frequency of occurrence of a stimulus vector **X** within the environment. Within this framework, we can view a broad class of neural network models as specific gradient ascent algorithms that maximize P(**X**), while some popular neural network learning algorithms are viewed as procedures that estimate the general form of P(**X**). In particular, the learning process is viewed as a procedure that constructs a P(**X**) such that P(**X**) obtains a local maximum for each class of vectors learned by the system.

Memory Recall is Maximizing a Probability Density Function

The Brain-State-in-a-Box (BSB) neural model (Anderson, Silverstein, Ritz, & Jones, 1977) is an abstract nervous system model that was designed to study various psychological phenomena. Information in this system is represented by an N-dimensional state vector that specifies a particular pattern of firing rates over a group of N neurons. The *i*th neuron in the system is modelled as a simple linear integrator possessing a maximum and minimum firing rate. The system operates by amplifying an incoming activation pattern (state vector) using positive feedback until many of the neurons in the system have obtained their maximum or minimum firing rates. The assumption that neurons possess maximum and minimum firing rates implies that the neural activation pattern over the set of N neurons is constrained to lie within an N-dimensional hypercube. More formally the BSB model is defined as follows:

$$x_i(k+1) = S[x_i(k) + \sum_j a_{ij}x_j(k)] \qquad (1)$$

where $x_i(k)$ is the *i*th element of the state vector **X** at discrete time interval k, and a_{ij} is a synaptic weight representing the synaptic efficacy between the *i*th and *j*th neurons in the system. The linearized sigmoidal function S[a] is defined as follows. S[a] = +1 for $a > +1$, S[a] = -1 for $a < -1$, and S[a] = a for $-1 \leq a \leq +1$.

Now let $\mathbf{X}_0$ be the initial value of the system state vector. Let E(**X**) be defined as:

$$E(\mathbf{X}) = (-1/2)\mathbf{X}^T\mathbf{A}\mathbf{X} \qquad (2)$$

where the *ij*th element of the matrix **A** is the synaptic weight a_{ij}. Golden (1986) has demonstrated that the BSB model is an algorithm that transforms $\mathbf{X}_0$ into a new vector **X**, located in the vicinity of $\mathbf{X}_0$, such that $E(\mathbf{X}) \leq E(\mathbf{X}_0)$ under fairly general conditions.

Now define P(**X**) as the probability of **X** and let the form of P(**X**) be given as follows:

$$P(\mathbf{X}) = ke^{-E(\mathbf{X})} \qquad (3)$$

where k is a constant chosen such that $\int P(\mathbf{X}) = 1$ and E(**X**) is defined in (2). The gradient of P(**X**) with respect to **X** is calculated as follows:

$$\mathbf{GRAD}[P(\mathbf{X})] = -\,\mathbf{GRAD}[E(\mathbf{X})]P(\mathbf{X}). \qquad (4)$$

Equation (4) states that an algorithm that moves along the path of steepest descent with respect to $E(\mathbf{X})$, is also moving along the path of steepest ascent with respect to $P(\mathbf{X})$. Moreover, when the state vector is "improbable" (i.e., $P(\mathbf{X})$ is small), the step size will be small. But when the state vector is "probable" (i.e., $P(\mathbf{X})$ is large), the step size will be large. Finally note that since $P(\mathbf{X})$ is a monotonically decreasing function of $E(\mathbf{X})$, a neural network model that minimizes $E(\mathbf{X})$ is also maximizing $P(\mathbf{X})$. In psychological terms, the BSB model is constructing a "more probable" interpretation of the initial state vector $\mathbf{X}_0$.

Learning is Estimating the Form of the Probability Density Function

In the BSB model, the synaptic weight between the *i*th and *j*th neurons in the system is specified by a real number, a_{ij}, that corresponds to the *ij*th element in the **A** matrix. The set of synaptic weights specify the parameters of the probability density function in (3) and therefore also specify the "knowledge base" of the system. To obtain a set of synaptic weights responsive to a particular set of training stimuli, the autoassociative Widrow-Hoff learning rule (Anderson, 1983) is used. More specifically, at each learning trial, a state vector is randomly selected from the set of training stimuli. Next, this training stimulus is used to update the current set of synaptic weights according to the synaptic weight updating rule:

$$a_{ij}(k+1) = a_{ij}(k) + \gamma\,(x_i - \sum_m a_{im}(k)x_m)x_j \qquad (5)$$

where $a_{ij}(k)$ is the value of the synaptic weight between the *i*th and *j*th neurons in the system at learning trial k, x_i is the *i*th element of the training stimulus vector, and γ is a positive learning constant. The on-diagonal elements, a_{ii}, are not updated.

Let **A** be a matrix of synaptic weights formed by the coefficients a_{ij}. Let **X** be the random vector associated with some unknown stationary probability distribution function in the environment. The autoassociative Widrow-Hoff learning rule can be shown to be searching for an **A** matrix that minimizes the expected value of the Euclidean distance between **AX** and **X** where the expectation is taken with respect to **X** (Widrow, 1971). Let **C** be a value of the random vector **X**. Golden (in preparation) has demonstrated that if **C** is a hypercube vertex, **A** is symmetric, and **AC** is in the same quadrant of the hypercube as **C**, then **C** is a strict local maximum of the density function. In conjunction with the observation that (2) is an energy or Liapunov function, this implies that a region about **C** exists such that any state initiated in that region must approach **C** as time increases.

Psychologically, these arguments simply indicate that the autoassociative Widrow-Hoff learning rule connects the neurons in the system such that the neural network *implicitly* assigns high probabilities to stimuli that have been taught to the system. These neural interconnections are then used by the BSB neural network to reconstruct "more probable" interpretations of less probable or novel state vector stimuli.

The Causal Schema neural network model

The Causal Schema (CS) neural model is a special type of production system specifically designed to model causal schemata. The model makes specific qualitative predictions regarding the pattern of errors made by people in recalling short, simple stories from memory. To illustrate the operation and behavior of the model, an experiment performed by Bower, Black, and Turner (1979) is described and then simulated using the CS neural model. Additional tests of the model are discussed by Golden (in preparation).

Bower et al. (1979) had college students learn a series of very short stories that were organized about routine event sequences or "scripts." Some of the stories were generated from the same script and were therefore very similar to one another (e.g., "visiting a doctor" and "visiting a dentist"), while other stories studied by the subjects were quite distinctive. After an intervening task, the subjects were given the titles of the stories as cues and requested to recall the actions that were mentioned in the story. Bower et al. found that "stated" script actions (i.e., actions explicitly mentioned within a story) were recalled more frequently than "unstated" script actions (i.e., actions implicitly mentioned), and "unstated" script actions were recalled more frequently than "other" types of script actions. In addition, as the similarity between two stories learned by a subject was increased, the number of "unstated" script actions recalled by the subjects increased and the number of "stated" script actions recalled by the subjects decreased (Table 1).

Constructing a Long-term Memory

The first step to modelling the Bower et al. (1979) experiment is the development of a long-term memory for the CS neural model. Such a memory was constructed in the following manner. Four distinct causal schemata associated with the event sequences "going to a restaurant," "going to a fast food restaurant," "going to a lecture," and "going to a doctor" were constructed. Next, two variations upon each of these four basic causal schemata were constructed. The resulting set of twelve schemata are then completely specified by a total of 107 causal relationships. A matrix of synaptic weights was then constructed from this stimulus set of causal relationships by training the system using the autoassociative Widrow-Hoff learning rule for 1000 learning trials. The resulting set of synaptic weights was defined as the model's long-term memory. In the simulations described here, five such matrices were generated using different random number seeds in an attempt to model the long-term memory structures of five college students.

Golden (in preparation) describes some computer simulations illustrating how this type of long-term memory system can be used to control behavior. In particular, an incomplete CR representing an initial situation (e.g., (S1,0,0)) is presented to the BSB model which reconstructs the action field of the CR. The effect of this action upon the environment results in a new situation (e.g., (S2,0,0)) that, in turn, can be used by the BSB model to reconstruct the second action in some action sequence.

Modelling the Learning of Short Stories

After the 1000 learning trials using the "long-term memory" stimulus set of 107 CRs were completed, the simulated "subjects" were trained with 24 "story sets." In particular, each subject was taught a single story set for an additional 100 learning trials and then tested. Each story set consisted of two similar stories derived from the same causal schema and one very distinctive story derived from another causal schema. A "story" simply consisted of a collection of five causal relationship state vectors that were implicitly linked together using the causal schema state vector encoding procedure that was described earlier. Note that the system's knowledge of stories is stored over the *same* set of synaptic weights as the system's long-term memory.

Modelling the Recall Process

Figure 3 illustrates the main flow of control when the CS neural model is requested to recall a story from memory. The BSB model is provided with an initial situation and action field (corresponding to the title of the story), and reconstructs the final situation field. The final situation field is then used to form the initial situation field of a new state vector. The action and final situation fields of this new state vector are filled with zeroes. The new state vector is

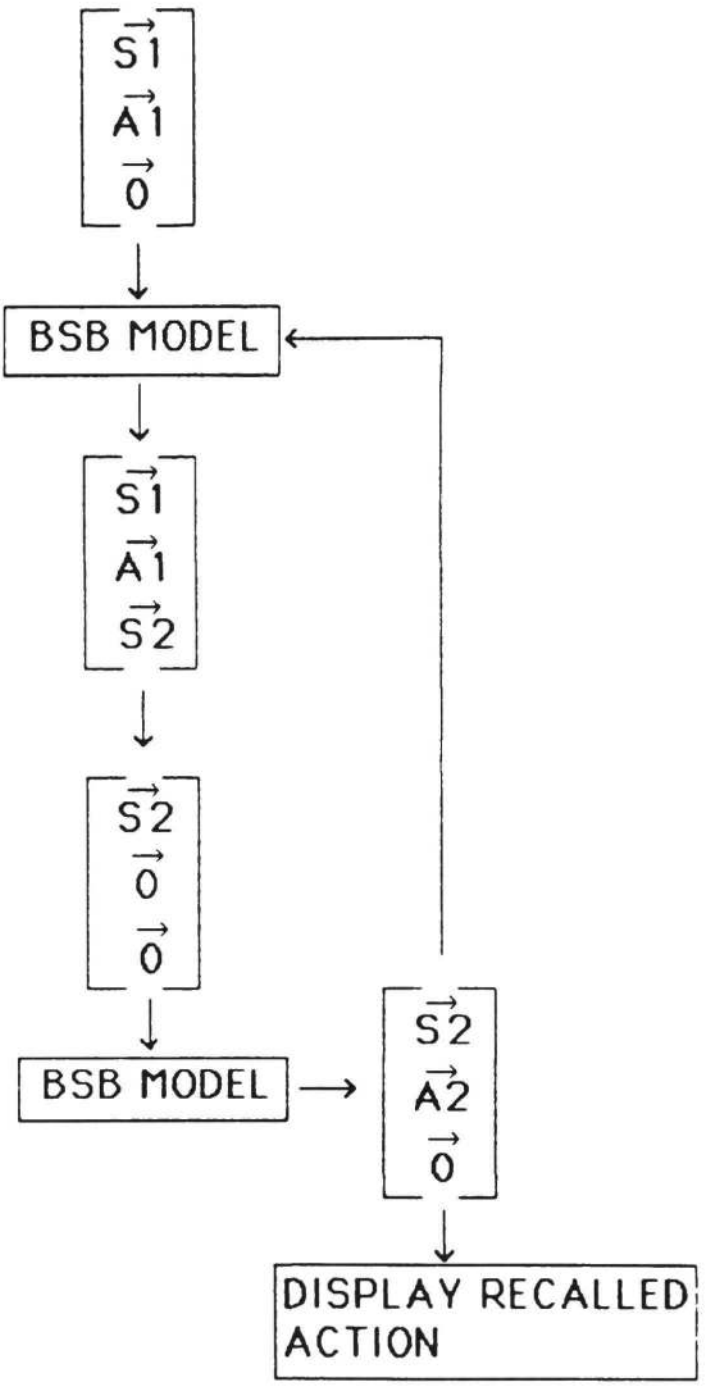

Figure 3. Flow of control during story recall. The CR representing the story title, (S1,A1,0), is transformed by the BSB model into (S1,A1,S2) thus reconstructing a final situation field for the partially specified (S1,A1,0). The final situation of (S1,A1,S2) is then used to form (S2,0,0) which is presented to the BSB model. The BSB model recalls an action, A2, from memory and this action is recorded by the experimenter. CR (S2,A2,0) may now be used as a memory cue to recall the next action in the story and the above cycle is repeated.

then submitted to the BSB model which reconstructs a new action. This new action is recorded as the first action recalled by the model, and the new action and new initial situation are used to initiate the cycle once again to recall the second action from memory.

Computer Simulation Results

Table 2 provides the results of the computer simulations which may be compared with the results obtained in the Bower et al. (1979) study. Like the human data in Table 1, "stated" actions are recalled more frequently than "unstated" actions which are recalled more frequently than "other" actions. In addition, as the number of related stories that are learned by the computer subjects increases, the number of stated actions recalled decreases and the number of unstated actions recalled increases. The interaction of "number of related stories" and "action type" was highly significant ($p < 0.01$) in the computer simulations treating either story sets or computer subjects as random factors.

Summary

A connectionist model of causal schemata in human memory has been described that makes specific qualitative predictions about experiments involving memory for written text. As an example, the performance of the model was compared with human subjects' performance in a specific psychological experiment. For this particular experiment, the CS model successfully captured the general qualitative characteristics of human recall memory for simple stories. In addition, a procedure for representing complex causal schemata as collections of neural activation patterns (state vectors) and a probabilistic interpretation of memory recall and learning in the BSB model were discussed.

Acknowledgements

This research was supported in part by a grant from the National Science Foundation to J. A. Anderson, administered by the Memory and Cognitive Processes section (Grant BNS-82-14728). I am grateful to David Cooper for suggesting that some analyses of connectionist models might be viewed within a probabilistic framework. I would also like to thank the members of the Brown University neural modelling group (particularly Jim Anderson and Mike Rossen) for their comments and advice.

References

Ackley, D. A., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147-169.

Anderson, J. A. (1983). Cognitive and psychological computation with neural models. *IEEE transactions on systems, man, and cybernetics, 5,* 799-815.

Anderson, J. A., Golden, R. M., & Murphy, G. L. (1986). *Concepts in distributed systems.* Paper presented at S.P.I.E. Advanced Institute Series Hybrid and Optical Computers, Leesburg, Virginia.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413-451.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177-220.

Golden, R. M. (1986). The "Brain-State-in-a-Box" neural model is a gradient descent algorithm. *Journal of Mathematical Psychology, 30*, 73-80.

Golden, R. M. (in preparation). *Modelling causal schemata in human memory: A connectionist approach.* Unpublished doctoral dissertation, Brown University, Providence, RI.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*, 2554-2558.

Hopfield, J. J. (1984). Neurons with graded response have collective properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA, 81*, 3088-3092.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension.* Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. v. 1: Foundations.* Cambridge, MA: Bransford Books/MIT Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Erlbaum.

Smolensky, P. (1984). The mathematical role of self-consistency in parallel computation. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society.* Boulder, Colorado.

Touretzky, D. S., & Hinton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the Ninth International Joint Conference on Artifical Intelligence, v. 1.*, Los Angeles, CA., pp. 238-243.

Widrow, B. (1971). Adaptive filters. In R. E. Kalman and N. DeClaris (Eds.), *Aspects of network and system theory.* New York: Holt, Rinehart, & Winston.

Table 1

Average number of actions recalled by human subjects
(adapted from Bower et al., 1979)

		Number of stated actions	Number of unstated actions	Number of other actions
Number of related stories	1	3.03	0.80	0.39
	2	2.27	1.26	0.35

Table 2

Average number of actions recalled by the CS neural model
(Computer simulation)

		Number of stated actions	Number of unstated actions	Number of other actions
Number of related stories	1	2.47	0.10	0.02
	2	1.62	0.40	0.00