Pathogenesis of Schizophrenic Delusions and Hallucinations: A Neural Model

Eytan Ruppin, MD, PhD *
School of Mathematics and School of Medicine
Tel Aviv University, Tel Aviv 69978, Israel

James A. Reggia, MD, PhD †
Department of Neurology
University of Maryland Hospital
22 South Greene St.
Baltimore, MD 21201

David Horn, PhD
School of Physics
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, Tel Aviv 69978, Israel

March 19, 1995

Abstract

We implement and study a computational model of Stevens' [1992] theory of the pathogenesis of schizophrenia. This theory hypothesizes that the onset of schizophrenia is associated with reactive synaptic regeneration occurring in brain regions receiving degenerating temporal lobe projections. Concentrating on one such area, the frontal cortex, we model a frontal module as an associative memory neural network whose input synapses represent incoming temporal projections. Modeling Stevens' hypothesized pathological synaptic changes in this framework results in adverse side effects reminiscent of hallucinations and delusions seen in schizophrenia: spontaneous, stimulus-independent retrieval of stored memories focused on just a few of the stored patterns. These could account for the occurrence of schizophrenic delusions and hallucinations without any apparent external trigger, and for their tendency to concentrate on a few central cognitive and perceptual themes. The model explains why schizophrenic positive symptoms tend to wane as the disease progresses, why delayed therapeutical intervention leads to a much slower response, and why delusions and hallucinations may persist for a long duration when they occur.

*This research has been supported in part by a Rothschild Fellowship to Dr. Ruppin, while he was a postdoctorate fellow at the University of Maryland. Correspondence should be addressed to Dr. Ruppin.

†Dr. Reggia is also with the Department of Computer-Science at the University of Maryland.

1 Introduction

Neural modeling research is currently a very active and growing scientific field with intense, multidisciplinary activity. The main emphasis in the past has been on the investigation of cognitive and neural functions in normal, healthy subjects. Recently, there has been a growing interest in the use of 'lesioned' neural models to investigate various brain pathologies and their cognitive and behavioral effects. To gain insight into how specific pathological neuroanatomical and neurophysiological changes can result in various clinical manifestations, the intact model's structural components may be lesioned, or its functional mechanisms may be disrupted. Recent published examples of such lesion studies include models of cortical plasticity following stroke [Armentrout *et al.*, 1994], memory impairment in Alzheimer's disease [Horn *et al.*, 1993; Ruppin & Reggia, 1995; Hasselmo, 1994], and cognitive and behavioral explorations of aphasia and acquired dyslexia [Reggia *et al.*, 1988; Dell, 1986; Hinton & Shallice, 1991].

In parallel to the computational investigations of neurological disorders, some pioneering steps in modeling schizophrenia have recently been undertaken. The two main directions have been the modeling of schizophrenic positive symptoms, and the modeling of various aspects of cognitive functions of schizophrenics. The first avenue, taken by Hoffman, has concentrated on modeling schizophrenic positive symptoms in the framework of an associative memory network [Hoffman, 1987; Hoffman & Dobscha, 1989]. This work has pointed to a possible link between the appearance of specific neurodegenrative changes and the emergence of 'parasitic foci', states in which a neural network's normal processing is disrupted and it is locked in dysfunctional patterns of activity. This result was followed by a series of clinical experiments that explored the possible role of such parasitic foci in the formation of schizophrenic positive symptoms [Hoffman & McGlashan, 1993]. Working in a second modeling framework, Cohen and Servan-Schreiber have provided a detailed computational account explaining how some schizophrenic functional deficits can arise from neuromodulatory effects of dopamine hypothesized in schizophrenia [Servan-Schrieber *et al.*, 1990; Cohen & Servan-Schreiber, 1992]. For a detailed review of modeling studies of neuropsychiatric disorders see [Reggia *et al.*, 1994; Ruppin, 1995].

While past studies have concentrated on investigating basic information processing disturbances that may be involved in the pathogenesis of schizophrenia, the goal of the present work is to examine the possibility of developing a neural model of a specific neurobiological theory of the pathogenesis of schizophrenia. We are primarily interested in investigating whether such an abstract theory, defined in general 'macroscopic' anatomical and cognitive terms, may be realized within the framework of a detailed (albeit simplified) neural model, and if so, what the possible insights are that may be gained from such a computational realization. More specifically, we have chosen to model a recent theory by [Stevens, 1992].

As summarized by Stevens [1992], the wealth of data gathered concerning the pathophysiology of schizophrenia suggests that there are atrophic changes in the hippocampus and parahippocampal areas in the brains of a significant number of schizophrenic patients, including neuronal loss and gliosis. On the other hand, neurochemical and morphometric studies testify to an expansion of various receptor binding sites and increased dendritic branching in the projection sites of medial temporal neurons, including a number of subcortical structures such as the nucleus accumbens, septum, thalamus, and cortical regions such as cingulate, prefrontal and medial frontal cortices. These findings have led Stevens to hypothesize that the onset of schizophrenia is associated with reactive anomalous sprouting and synaptic reorganization taking place in the projection sites of dystrophic medial temporal neurons.

To study the possible functional implications of Stevens' hypothesis, we concentrate in this paper on a single frontal module receiving temporal projections. Even though Stevens' hypothesis involves changes occurring in numerous cortical and subcortical structures, we think it is pertinent at this basic modeling stage to focus on a simple, canonical, computational model of Stevens' hypothesis, that, while including only a small subset of the brain structures involved, still encompasses the primary synaptic changes whose effects we want to study. The frontal cortex was chosen to be at the focus of our study primarily because, in addition to Stevens' theory, there is a considerable amount of data testifying for frontal lobe involvement in schizophrenia (see [Weinberger, 1991] for a comprehensive review). Moreover, this data suggests that prefrontal hypo-metabolism in schizophrenia may be a secondary effect relating to the function of abberant afferent temporal projections [Weinberger, 1991]: recent *in vivo* metabolic data indicate that prefrontal cortex and hippocampus are functionally coupled during memory tasks [Friedman *et al.*, 1990], and that sharing information via numerous temporo-frontal connections is an important aspect of prefrontal functioning [Golman-Rakic, 1987]. Over all, it seems that the disruption of temporo-frontal connections and reactive frontal sprouting may have an important functional role in the pathogenesis of schizophrenia. From an anatomical perspective, there are

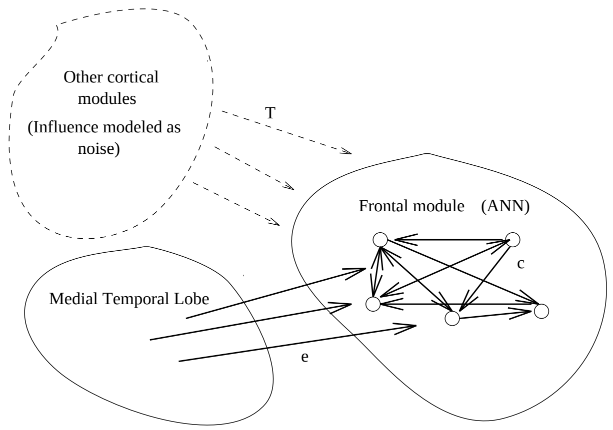

Figure 1: A schematic illustration of the model. Each frontal module is modeled as an attractor neural network whose neurons receive inputs via three kinds of connections: internal connections from other frontal neurons, external connections from temporal lobe neurons, and diffuse external connections from other cortical modules, the latter modeled as noise.

both direct routes to frontal lobes from medial temporal lobe structures, and indirect routes via the thalamus and nucleus accumbens [Nauta & Domesick, 1982]. For simplicity, we currently focus on studying the possible functional effects of damage in the direct route, which may also approximate the cascading effect of weakening temporo-subcortical connections on indirect temporo-frontal routes.

The frontal cortex is a plausible site of working memory in the brain [Goldman-Rakic, 1991], and, together with other areas of association cortex, may also be an important storage site of long-term, associative, content-addressable memory. Adopting the latter stance, we assume that memory retrieval from the frontal cortex is invoked by the firing of incoming temporal projections. This assumption is motivated by the idea that temporal structures have an important role in establishing long-term memory in the neocortex and in the retrieval of facts and events (e.g., [Heit *et al.*, 1988; Squire, 1992; Alvarez & Squire, 1994]), and by the strong functional links between temporal and frontal areas during memory processes. As illustrated in Figure 1, a frontal module is modeled as an associative memory neural network, storing memorized patterns in a Hebbian, internal, synaptic matrix. Such a frontal module represents a macro-columnar unit that has been suggested as a basic functional building block of the neocortex [Goldman & Nauta, 1977; Mountcastle, 1979; Eccles, 1981]. It receives specific memory-cue inputs from temporal projections, and is interconnected with other areas via diffuse inter-modular connections. In accordance with Stevens' theory, when schizophrenic pathological processes occur, the degeneration of temporal projections is modeled by decreasing the strength of the incoming temporo-frontal input fibers. Reactive frontal synaptic regeneration is modeled by increasing the strength of the internal frontal connections, and the expansion of diffuse external projections is modeled by increasing the noise level effecting the network dynamics. The latter is obviously a crude simplification of the possible role of the inter-columnar cortical projections, but is partially justified since cue input patterns are assumed to arise specifically from temporal lobe projections, so that the effect of other cortical inputs during a memory retrieval episode may be viewed as noise.

The current work is a first attempt to construct a neural model that examines the role of changes in synaptic connections projecting from one specific cortical region to another in the pathogenesis of schizophrenia. Section 2 describes the neural model used. The numerical experiments conducted within this framework are presented in Section 3. The relevance of our results to Stevens' theory and to the pathogenesis of schizophrenia are discussed in Section 4. A more technical description of the model's details, and a computational analysis testifying to the broad class of associative memory neural models to which our results apply, are provided in separate appendices.

2 Methods

Our model is based on viewing the frontal cortex as composed of multiple associative memory modules. We concentrate on the changes that occur in a single cortical module, which is modeled as an attractor neural network. An *attractor neural network* [Hopfield, 1982; Amit, 1989] is an assembly of formal neurons connected recurrently by synapses (see Figure 1). The network's state, that is, the collective firing state of its neurons, is repeatedly updated; when a neuron fires, its output, weighted by synaptic strengths, is communicated to the neurons to which it is connected. This spreading activity serves as input to those neighboring neurons, and may, in turn, trigger them to fire. By using specific learning rules that govern the way synaptic strengths in the network are established, a specific set of input patterns can be memorized, i.e., made to be 'attractors' of the network dynamics. The term 'attractor' here means that, if a pattern which is sufficiently similar to one of the stored memory patterns is presented as input to the network, the network's state will gradually evolve until it converges to the state representing that memory pattern.

Such a network may therefore be regarded as an associative memory system.

In attractor neural networks, stored memories are not represented locally (at specific neurons of the network), but their corresponding representations are distributed; many neurons participate in a given memorized pattern, and a particular neuron participates in several different patterns. Representing stored memories as attractors corresponds to our intuitive notion of the persistence of cognitive concepts along some temporal span. It also is supported by biological findings of delayed, post-stimulus, sustained activity in memory-related tasks, both in the temporal [Fuster & Jervey, 1982; Miyashita & Chang, 1988] and frontal [Wilson *et al.*, 1993] cortices. These experiments show that there are cortical neurons which continue to fire at an increased rate for a few seconds even after the external stimulus that originally triggered their response has been removed. These persistent firing reverberations have been observed in localized modules, each about 1 mm in diameter, and are not a single neuron property but reflects a collective behavior [Miyashita & Chang, 1988; Sakay & Miyashita, 1990].

We use a biologically-motivated variant of attractor neural network model, proposed by Tsodyks and Feigelman [Tsodyks & Feigel'man, 1988]. The network is composed of N neurons, where each neuron may be in either an active (firing) or passive (quiescent) state, respectively. Each neuron's state is updated stochastically, in accordance with its synaptic inputs (membrane potential), i.e., the signals it receives from the active neurons in the network and from external, inter-modular, sources. If a neuron's current membrane potential is significantly higher than its firing threshold, that neuron's state will be active, and otherwise it will remain silent, with high probability (see *Appendix A* for technical details). An active neuron, in turn, influences the membrane potential of other neurons by transmitting a spike via its outgoing synaptic connections. Its effect on other neurons depends of the sign and magnitude of these synapses.

As illustrated in Figure 1, the neurons receive three kinds of connections: 1. *External input connections*, representing temporal projections via which external input patterns are presented to the network. The degeneration of temporal projections is modeled by a uniform decrease in their strength parameter e. 2. *Internal connections*, which store M memorized patterns and represent the intra-modular frontal connections. The magnitude of these connections are determined via a Hebbian, activity-dependent learning rule, which strengthens the synaptic connections between neurons that are firing together and decreases the connection strength (i.e., lowers the synaptic weights) between neurons whose activation state is uncorrelated. Frontal synaptic regeneration is modeled by increasing their strength parameter c. 3. *External diffuse connections*, representing 'non-specific' inter-modular connections. The latters' effect on the network dynamics is expressed via the noise level T, which is increased to model diffuse synaptic regeneration.

In our computer simulation experiments, the functioning of the network is examined in two scenarios:

1. In the *stimulus-dependent retrieval* scenario, a stored memory pattern is presented as an input cue to the network via the external synaptic projections, the network state evolves until it converges to a stable state. In its normal, premorbid state, the network will converge to a state highly similar to the cued memory, denoting successful retrieval. However, if the external input projections are severely weakened the network will either wander around in an autonomous state of random, low-activity, or converge to a *mixed* state where it does not attain high similarity with any of the stored memory patterns.

2. In addition to investigating the network's activity in response to the presentation of an external input cue, we examine its behavior in the absence of any specific stimulus input cue, denoted as *spontaneous retrieval*. In this scenario, the external input synapses are non-active and the outcome of the network's behavior in each trial depends only on its random initial state and the dynamics governed by the internal synaptic connections. In its premorbid state, the network continues to wander around in a state of random low baseline activity. However, as we shall show, following synaptic compensatory changes that preserve memory retrieval by strengthening the magnitude c of the internal synaptic connections and by increasing the noise level T, the network - *without being cued* - may converge onto a stored memory state, resulting in a pathological, autonomous activation of patterns memorized by network.

The memory performance of the network is quantified by measuring its retrieval accuracy. In each trial, the network is initiated with some initial random pattern of activity and its behavior is traced. After the network has converged to a stable state, or after a certain amount of time has elapsed if the network does not converge to a stable state, we measure the similarity between the network's state of activation and the cued memory pattern, on a scale from 0 to 1. Similarity level 1 denotes perfect retrieval of the cued memory pattern, and a memory pattern is considered to be retrieved if the network converges to a stable state which has similarity greater than 0.9 with it. In a given experiment, that is, with some fixed levels of synaptic magnitudes and noise, the performance of the network is assessed by averaging the similarity level achieved over 100 trials.

Numerical experiments are performed either in the stimulus-dependent or the spontaneous-retrieval scenarios, that is, with or without the presence of an external input cue. The networks used have either $N = 400$ or $N = 800$ neurons, storing $M = 20$ (or 40) memory patterns, respectively. (The simulations involving activity-dependent changes required larger networks to avoid 'finite size' effects and were hence performed in a network of $N = 800$ neurons). In the initial, premorbid state, the parameter value determining the external input synaptic strength is $e = 0.035$, the internal synaptic strength parameter is $c = 1$, and the noise level (the external diffuse synaptic input) is $T = 0.009$. These parameter values ensure that in its intact premorbid state the average similarity attained is almost 1, i.e., the retrieval performance of the network is near perfect.

3 Experiments and Results

We turn now to simulations examining the behavior of the model network under variations of synaptic strength and noise level. Simulating the changes occurring in schizophrenia in accordance with Stevens' theory, the external input synapses' magnitude (e) are weakened, and the internal synapses magnitude (c) and noise level (T) are increased.

First, in the experiments described in Section 3.1, we demonstrate that either strengthening internal synapses or increasing noise levels results in the emergence of spontaneous activation of stored patterns, which are pathologically retrieved in an autonomous manner without being cued. However, as we shall show, the latter regenerative synaptic changes (i.e., increasing the internal synaptic strength or the noise level) subserve a functional role, enabling the maintenance of memory retrieval capacities even though external input synapses are weakened. During this phase, the internal synapses are strengthened in a simple homogeneous manner, increasing their magnitude by a common fixed factor.

Second, in the experiments described in Section 3.2, we adopt the assumption that the increase in the magnitude of the internal synapses storing the memorized patterns involves an additional activity-dependent Hebbian mechanism, similar to the learning rule via which the memorized patterns were initially stored. Thus, we assume that the primary pathologic process proposed by Stevens (the degeneration of external synapses) triggers a compensatory increase in internal memory-bearing synapses, and that the latter has combined non-activity-dependent and activity-dependent components. Incorporating activity-dependent synaptic changes into the dynamics of the model yields an interesting result: the distribution of the pathologically spontaneously retrieved patterns becomes concentrated on just a very few patterns. An investigation of the combined effect of activity-dependent changes and spontaneous retrieval on the retrieval properties of the network reveals some additional findings, which bear an interesting resemblance to some of the characteristic features of schizophrenic delusions and hallucinations.

3.1 The Emergence of Spontaneous Retrieval

After weakening the external synapses (to the level $e = 0.015$), we examine the behavior of the network in the spontaneous-retrieval scenario, without any input cues. That is, on each simulation trial the network is initiated in a random low-activity state, the input external field is shut off, and the network state evolves as described in the Methods Section. Recall that in its premorbid state, the network will remain in a low-activity firing state and will not retrieve any stored pattern in the absence of appropriate input cues. However, as shown in Figure 2, either synaptic strengthening or increased noise levels may result in spontaneous, erroneous retrieval of non-cued memory patterns: Beyond some critical level of increase in either the internal synaptic strength or the noise level, the network begins to frequently retrieve memory patterns although it does not receive any external input!

Schizophrenic symptomatology involves complicated cognitive and perceptual phenomena, whose description certainly requires much more elaborate representations than a simple neural model of associative memory. Yet, whatever their neural realization may be, schizophrenic delusions and hallucinations frequently appear in the absence of any apparent external trigger. It therefore seems plausible that the emergence of spontaneous activation of stored patterns described above is likely to be an essential element in their pathogenesis. It should also be noted that when spontaneous retrieval emerges, the network may spontaneously converge at times to non-stored patterns, which are a 'mixture' (i.e., have some similarity) of a few memory patterns [Horn & Ruppin, 1995a]. Such retrieval of mixed patterns my play part in explaining the generation of more complex forms of schizophrenic delusions and hallucinations, involving abnormal condensation of thoughts and imaginings.

What then may be the possible computational role of increased internal synaptic strength and noise level, modeling the regenerative synaptic changes taking place in accordance with

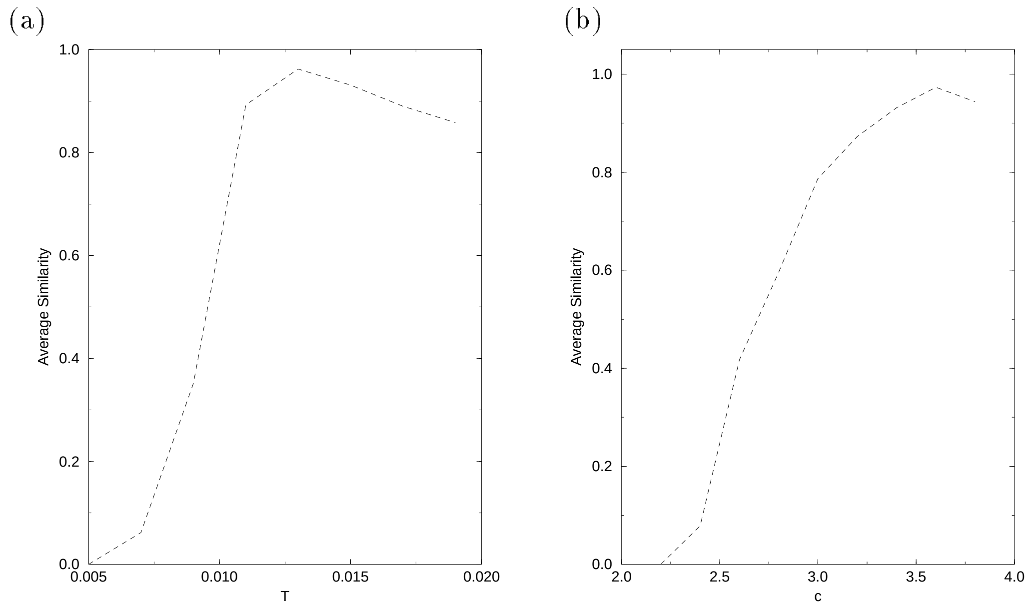

Figure 2: Spontaneous retrieval, measured as the highest final similarity achieved with any of the stored memory patterns, emerging in a network with decreased external input strength ($e = 0.015$) and increased noise or internal synaptic strength. (a) Spontaneous retrieval as a function of the noise level T. $c = 1$. (b) Spontaneous retrieval as a function of internal synaptic compensation factor c. $T = 0.009$.

Stevens' hypothesis? To answer this question, we have examined the retrieval performance of the network in the stimulus-dependent scenario. As in the spontaneous-retrieval scenario, in each trial the network's initial state is random, but now the network states evolves in the presence of an input memory pattern cue, which is applied to the network via the (albeit weakened) external input synapses. The network's retrieval performance is quantified by measuring the average similarity between the cued input patterns and the network's response (over a hundred trials, in each set of synaptic parameter values defining a given network).

Figure 3a displays simulation results showing that an increase in the noise level T can compensate for the deterioration of memory retrieval due to a decrease in the external input e. For fixed T, performance decreases rapidly as the external input strength e is decreased. However, if the decrease in e is not too large, an increase in T restores stimulus-dependent retrieval performance. The first three curves are qualitatively similar, characterized by a peak of the retrieval performance at some e-dependent optimal level of noise. Eventually, at very low external input strength levels retrieval is lost. Similarly, as shown in Figure 3b, an increase in the internal synaptic strength c may compensate for decreased external

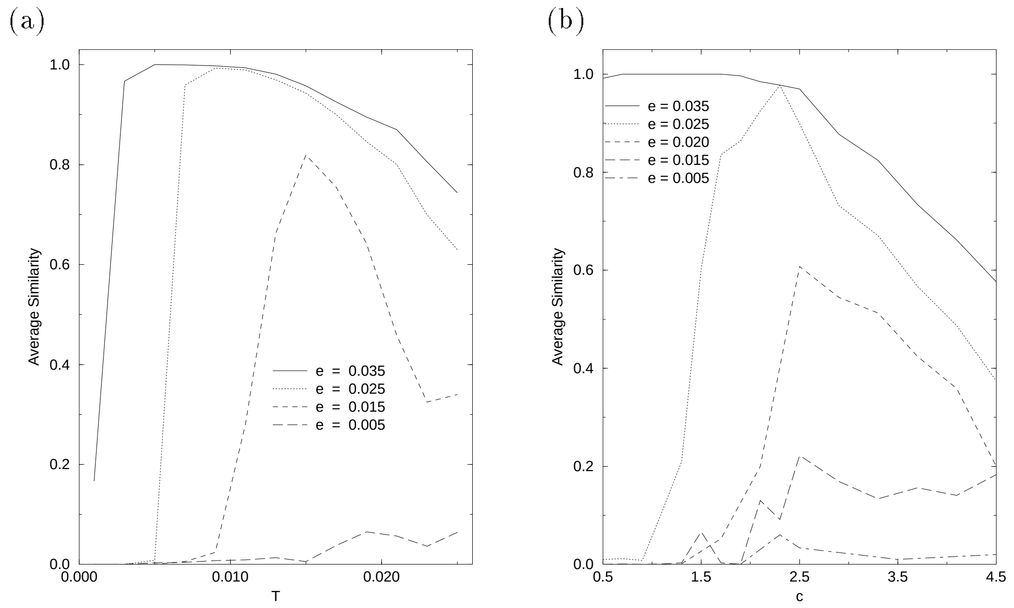

Figure 3: (a) Stimulus-dependent retrieval performance, measured by the average final similarity, as a function of the noise level T. Each curve displays this relation at a different magnitude of external input projections e ($c = 1$). (b) Stimulus-dependent retrieval performance as a function of the internal synaptic strength c ($T = 0.009$).

input strength. As e is decreased the best possible performance is achieved with increasing c values. The combined compensatory potential of internal synaptic strengthening and increased noise is synergistic, as high stimulus-dependent retrieval performance is achieved already at a fairly low increase of synaptic and noise levels. We see that these expansive synaptic changes, which represent the regenerative changes assumed by Stevens, do have a beneficial computational role in maintaining memory capacities in face of the weakened external input synapses (representing degenerated temporal projections).

3.2 Biased Spontaneous Retrieval

We now turn to study the effects of incorporating Hebbian, activity-dependent synaptic changes as part of the internal synaptic regeneration. This investigation is motivated by recent findings that increased dopaminergic activity may enhance Hebbian-like activity-dependent synaptic changes (as discussed in detail in Section 4), and that the density of NMDA receptors is increased in cortical areas of schizophrenics [Javitt & Zukin, 1993]. It should be emphasized that the activity-dependent synaptic modification mechanism we employ is essentially the same Hebbian activity-dependent mechanism that is employed for the storage of memorized patterns during a learning episode (*Appendix A*). We thus assume that the synaptic modification rate, denoted γ, has significant magnitude during the early 'childhood' plastic period. It later decreases to near zero levels, maintained throughout 'adulthood', and is adversely increased with the synaptic regenerative processes associated with the onset of schizophrenia. As we shall show, the same Hebbian synaptic modification mechanism that underlies normal memory storage can lead to increasing damage to the associative memory stores when employed in pathological conditions leading to spontaneous retrieval.

In the following simulations we examine the network's behavior after the pathological changes hypothesized by Stevens have taken place, including degenerative loss of temporal projections and regenerative compensatory synaptic changes ($e = 0.015$, $c = 1.5$, $T = 0.017$). First, we trace the behavior of the network in the *spontaneous retrieval* mode during many trials, each starting form a different initial random state. Due to the compensatory synaptic changes, some of the memorized patterns are now spontaneously retrieved. Because of the incorporation of activity-dependent synaptic changes ($\gamma = 0.0025$), the synaptic matrix does not remain fixed any more. As the synaptic matrix is retained from trial to trial, it gradually evolves as spontaneously-generated patterns of the activity are engraved into it. This, in turn, effects the future dynamic behavior of the network.

Figure 4 traces the distribution of the memory patterns the network has spontaneously converged to during the last one hundred trials preceding the 200'th trial, the 500'th trial and the 800'th trial. The total frequency of convergence to memory patterns increases as time evolves (i.e., in later trials): from 0.46 after 200 trials to 0.68 after 500 trials to 0.98 after 800 trials. As is evident in Figure 4, the distribution of the memory patterns spontaneously retrieved tends to concentrate on a single memory pattern as more trials occur. Although the synaptic matrix was initially non-biased, small, random correlations between the network's initial states and a few of the memory patterns are sufficient to overwhelmingly enhance their retrieval. We therefore see that biased retrieval is formed, and out of the many patterns stored in the network only very few are actually spontaneously retrieved. This formation of a biased spontaneous retrieval distribution when Hebbian activity-dependent synaptic changes accompany the reactive internal synaptic strengthening is in accord with the common finding that delusions and hallucinations tend to concentrate upon a limited set of recurring cognitive and perceptual themes (e.g., [Hoffman, 1986; Chaturvedi & Sinha, 1990]),

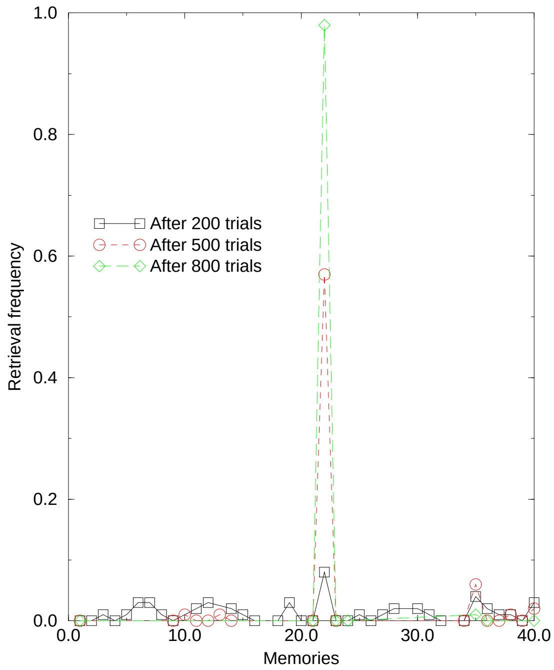

Figure 4: The distribution of memory patterns spontaneously retrieved. The x-axis enumerates the memories stored, and the y-axis denotes the retrieval frequency of each memory.

The highly peaked, biased distribution of memory retrieval observed in the spontaneous retrieval mode is maintained for a few hundred additional trials, until memory retrieval sharply collapses to near zero as a global *mixed-state* attractor is formed. Such a mixed attractor state has considerable overlap with a few memory patterns, but does not have very high overlap with any memorized pattern, and is thus considered not to stand for a well-defined cognitive or perceptual item [Amit, 1989]. Once a mixed attractor is formed the network converges to it on each trial, that is, this attractor completely dominates the activity of the network. It is an end state of the Hebbian, activity-dependent evolution of the network; extensive simulations show that once an mixed attractor is reached, the network will remain in its vicinity practically forever.

In the previous simulation all memory patterns were stored with equal strength in the synaptic matrix. Even so, the small correlations that exist between the randomly generated memory patterns are sufficient to generate a single-peaked retrieval distribution in the spontaneous retrieval scenario. To examine the effect of an initial bias in memory storage, we randomly pick one of the memories (say #1) and store it with strength (weighting) $k > 1$ times more that all other memories. In the absence of other significant biases, an initial bias as small as $k = 1.1$ is sufficient to markedly shift the retrieval distribution towards the biased memory (#1). Thus, when spontaneous retrieval emerges the network has a strong tendency to markedly amplify pre-existing biases in its internal synaptic matrix.

Next we turn to examine the distribution of memories retrieved when the network operates in the *stimulus-dependent retrieval* scenario. The network is similar to the one employed above, but now, on each trial, a memory pattern is randomly chosen and presented as an external input to the network. As illustrated in Figure 5, the resulting distribution of retrieved memories is not concentrated around any memory pattern. After about 500 trials a global mixed-state attractor is formed, and the network looses its capacity to perform stimulus-dependent retrieval. Moreover, even when the internal synaptic memory matrix has an initial bias, the retrieval distribution obtained in the stimulus-dependent scenario will remain dispersed (this was found to hold up to levels of initial bias of $k = 2.5$). Thus, the external input, which is homogeneously distributed among the memorized patterns, *counter-acts the effect of the biased synaptic matrix* and impedes the formation of a retrieval distribution concentrated on a single pattern. The retrieval performance is preserved until late stages in the evolution of the synaptic matrix, when a mixed attractor is formed. Hence, while synaptic regenerative changes may lead to the emergence of spontaneous retrieval in frontal cortical modules, it is the denervation of external input projections that actually makes the frontal networks susceptible to the formation of a biased spontaneous retrieval distribution.

In another simulation experiment, the synaptic matrix is generated in a homogeneous, unbiased manner. However, the external input distribution applied during a stimulus-dependent retrieval epoch is now no more homogeneous, but is strongly biased, i.e., one memory pattern is presented to the network as an input cue $k > 1$ times more than any other. Remarkably, we find that in these conditions the retrieval distribution remains homogeneously distributed up to relatively large levels of bias of magnitude $k \approx 3.5$. These studies teach us that the mechanisms leading to the formation of biased spontaneous retrieval are primarily geared at amplifying any initial bias in the premorbid synaptic matrix, and are less sensitive to biases in the current input stream.

Finally, it should be noted that as activity-dependent synaptic changes take place, the absolute magnitude of the synapses constantly increases (see *Appendix B*). This constant synaptic weight increase, together with the notion that biological synapses probably cannot increase their effectiveness in an unlimited manner, obviously raises the question of the net-

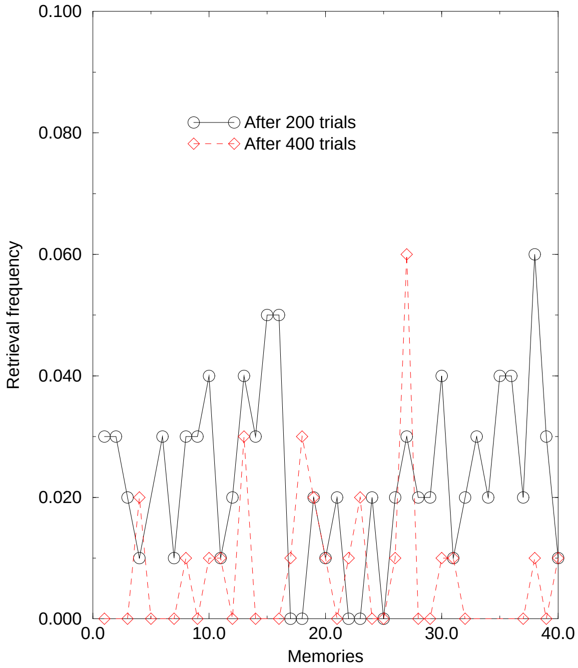

Figure 5: The distribution of stimulus-dependent retrieval of memories.

work's behavior in the presence of a bound on the absolute synaptic magnitude. Introducing a bound on the absolute magnitude of the weights of the internal synaptic matrix (of value 2.5), we have rerun the simulations described above, both in the stimulus-dependent and in the spontaneous retrieval scenarios. The retrieval distribution obtained is qualitatively similar to that observed without synaptic bounds. In its end-state, the network is 'trapped' into a mixed attractor state where all of its synaptic weight values are fluctuating near the synaptic bounds.

4 Discussion

We have shown that while preserving memory performance, compensatory synaptic regenerative changes modeling those proposed in Stevens' theory of schizophrenia may lead to adverse, spontaneous activation of stored patterns. When spontaneous retrieval emerges, the incorporation of Hebbian activity-dependent synaptic changes leads to a retrieval distribution that is not homogeneously dispersed, but is strongly dominated by a single memory pattern. A small initial bias in the memory matrix towards one of the stored patterns is sufficient to lead to its dominance in the spontaneous retrieval scenario. However, when a stream of external input stimuli arrives at the network in the stimulus-retrieval scenario, many of the stored patterns are retrieved and a more homogeneous retrieval distribution is obtained. Moreover, the distribution remains essentially homogeneous even when some external input patterns are presented as cues to the network more frequently than others. To summarize, a biased retrieval distribution results from the amplification of initial biases in the synaptic memory matrix during periods of spontaneous retrieval, while stimulus-dependent retrieval epochs tend to work against the formation of such a distribution.

The main conclusion of this study is that the formation of biased, spontaneous retrieval requires the *concomitant occurrence* of both degenerative changes in the external input fibers, and regenerative Hebbian changes in the intra-modular synaptic connections. Both kinds of synaptic changes are the main microscopic pathological processes that take part in the pathogenesis of schizophrenia in accordance with Stevens' hypothesis [1992]. Obviously, our study is based on a grossly simplified model of the relevant neural circuitry. Yet, in spite of that, our results testify to the plausibility of Stevens' theory by showing that the realization within a neural model of the hypothesized pathological synaptic changes leads to significant changes in the macroscopic behavior of the network, which share some of the characteristics of schizophrenic positive symptoms.

It should be emphasized that the emergence of spontaneous, biased retrieval in the model cannot occur if either external synaptic weakening (i.e., temporal lobe degeneration) or internal synaptic strengthening (i.e., frontal regeneration) occur separately. It is only when these synaptic changes occur in conjunction, in a certain given range of synaptic synaptic strengths, that autonomous pathological activation of stored patterns emerges. In contrast to what one may intuitively think, there are not many changes in the network parameters that may result in these pathological manifestations, while preserving memory retrieval. To begin with, analytic considerations show that the emergence of spontaneous retrieval requires considerable strengthening of internal or diffuse external synapses [Horn & Ruppin, 1995a]. In addition, as demonstrated in this paper, a spontaneous retrieval distribution will only become biased if the internal synaptic strengthening includes an activity-dependent component, and if the frequency of successful stimulus-dependent retrieval is markedly reduced. Moreover, even when spontaneous retrieval emerges and synaptic activity-dependent changes ensue, a biased retrieval distribution of memory patterns may not always form. Depending on various network parameters (e.g., if the rate of activity-dependent changes γ is too high), the network may almost instantly converge repeatedly into a mixed-attractor, without going into a phase of biased memory retrieval. In summary, spontaneous, biased memory retrieval occurs only when certain specific changes occur in the network. However, the range of synaptic changes resulting in biased memory activation is sufficiently broad to make these synaptic changes a plausible cause of the latter pathological abberations. In Appendix B. we show that this conclusion is true not only with respect to the specific network model used in this work, but also for a fairly broad family of associative memory models, where compensatory synaptic strengthening has an overall excitatory effect on the network activity.

The notion that biases in the synaptic memory matrix tend to amplify themselves into a dominant pathological attractor if activity-dependent changes occur at a sufficient rate, naturally raises the question of how the brain normally resists generating such pathological attractors in its healthy state. We believe that there are several defense mechanisms that keep this pathological tendency in check in the normal brain. First and foremost, note that the propensity of the network to amplify initial biases in its synaptic matrix is likely to manifest itself only if pathological, spontaneous (initially non-biased) retrieval emerges. As long as spontaneous retrieval is overruled by stimulus-dependent retrieval, the synaptic matrix will primarily reflect the (approximately) homogeneous distribution of memory patterns cued in the stimulus-dependent retrieval mode and suppress any bias formation, as we have demonstrated in the previous section. Second, initial biases are likely to be amplified only if the rate of activity-dependent changes is sufficiently high; we have found that only if $\gamma > 0.001$ does a biased retrieval distribution evolve. Accordingly, it is possible that the synaptic matrix becomes more susceptible to bias amplification in the acute phases of schizophrenia due to a dopaminergic induced increase in the rate of activity-dependent changes (discussed further below). Third, there may be local activity-dependent 'synaptic maintenance' mechanisms that in normal conditions can 'sense' the formation of biases in the memory matrix and counteract them [Horn & Ruppin, 1995b].

Our results hence provide considerable support for the plausibility of Stevens' ideas from a computational point of view, but it should be acknowledged that Stevens theory remains a hypothesis. Obviously, there is still considerable uncertainty concerning the pathological changes underlying schizophrenic symptomatology. On one hand, a broad range of findings support Stevens' proposal that synaptic regeneration occurs in target areas of degenerating temporal projections [Stevens, 1992] (including, e.g., increased glutamate uptake sites, expansion of NMDA biding sites, increased dendritic branching of pyramidal cells and increased levels of synaptophysinelike proteins). However, other studies suggest that a reduction of frontal connectivity (due to both reduced production of new dendritic-axonal connections, and their excessive pruning) may be typical of many schizophrenic patients [Petegrew *et al.*, 1991; Petegrew *et al.*, 1993], and that synaptic glutamate is reduced in prefrontal regions in schizophrenia [Sherman *et al.*, 1991]. It is possible that while on a large scale there is an overall reduction of synapses, there are local 'islands' of excess synaptic regeneration which can result in the emergence of spontaneous retrieval in a few cortical modules.

Interesting additional support to Stevens' idea that a disconnection between temporal and frontal systems plays an important role in the pathogenesis of schizophrenia comes from a metachromatic leukodystrophy model of schizophrenia [Hyde *et al.*, 1992]: In a number of cases with a rare adult type of metachromatic leukodystrophy, the disease initially presents itself with schizophrenia-like psychosis. These patients may be psychiatrically treated for years before neurological symptoms become manifest [Monowitz *et al.*, 1978]. As noted in [Weinberger, 1991; Hyde *et al.*, 1992], the pathological changes in this disease affect primarily white matter and glia, while frontal and temporal neurons are generally spared. Thus, in support of Stevens' theory, it seems that psychotic symptoms may arise because of synaptic disconnection, in the absence of significant neuronal damage.

Despite a number of suggestive findings, there is currently no proof that a global abnormality of neuro-transmission is a primary feature of schizophrenia [Mesulam, 1990; Williamson, 1993; Carpenter & Buchanan, 1994]. Modeling Stevens' theory, we have focused on neuroanatomical synaptic changes, without referring to any specific neurotransmitter. On the other hand, delusions and hallucinations are believed to be responsive to dopaminergic blocking agents. While it is possible that dopaminergic agents influence the dynamic behavior of the network by modulating the neuronal firing function [Servan-Schrieber *et al.*, 1990; Cohen & Servan-Schreiber, 1992], our model raises the possibility that synaptic activity-dependent modifications are enhanced by increased dopaminergic activity, superimposed on the non-Hebbian synaptic compensatory changes. That is, at least part of the therapeutic effect of dopaminergic-blocking agents in the reduction of schizophrenic positive symptoms may be due to the attenuation of Hebbian, activity-dependent synaptic changes. Indeed, recent data pertaining to synaptic long-term potentiation and long-term depression may lend support to this hypothesis, suggesting that dopaminergic influences during and immediately after tetanization contribute to the induction of postsynaptic mechanisms sub-serving a late, long-lasting maintenance of synaptic potentiation [Frey *et al.*, 1990]. In addition, haloperidol can block both the induction and expression of amphetamine-induced sensitization, which may be a behavioral manifestation of long-term potentiation [Karler *et al.*, 1991]. Finally, antagonists of either D1 or D2 dopamine receptors can block long-term depression, and in dopamine-depleted slices long-term depression can be restored by applying exogenous dopamine [Calabresi *et al.*, 1992].

In addition to showing that the synaptic changes occurring in accordance with Stevens' theory may lead to the spontaneous activation of a small set of memorized patterns, our results may account for a few characteristics of schizophrenic positive symptoms:

- First, note that the emergence of spontaneous, non-homogeneous retrieval is a self-limiting phenomenon; eventually, a global mixed-state attractor is formed which does not have high similarity with any of the stored patterns and is hence meaningless. Such a meaningless cognitive pattern, or alternatively viewed, the accumulating loss of memorized patterns' attractors, may contribute to the emergence of deficit, negative symptoms, as discussed by [Globus & Arpaia, 1994; Hoffman & McGlashan, 1993]. This parallels the clinical observation that as schizophrenia progresses positive symptoms tend to wane, while the negative symptoms are enhanced [Kaplan & Sadock, 1991; Gray *et al.*, 1991; Carpenter & Buchanan, 1994]. Of course, if the sprouting of synaptic regenerative changes ends before a global mixed attractor is formed, the network could remain 'frozen' in a state dominated by biased memory retrieval and positive symptoms would continue. The global mixed attractor discussed above should also be differentiated from the mixed attractors the network may converge to at an early 'evolutionary' phase, when spontaneous retrieval emerges and activity-dependent changes haven't yet made their mark (see Section 3.1). While the latter mixed states generally have significant overlap with a few memories and thus may be deemed meaningful, the global mixed attractor typically has only negligible overlap with any of the stored memories, and is thus considered to be meaningless.

- Second, when the network converges to a memory pattern that dominates the output in the spontaneous-retrieval scenario, it has an increased tendency to remain in this state for a much longer time than in its normal functioning state (see *Appendix B*). This is in accordance with the persistent characteristic of positive symptoms, which may endure for long time periods.

- Third, as more and more spontaneous retrieval trials occur, the frequency of spontaneous retrieval increases, until some point is reached where it sharply declines. In accordance, a similar pattern should be observed concerning the frequency of positive symptoms in schizophrenics as the disease progresses. In this regard it is worth noting that while early treatment in young psychotic adults leads to early response within days, late, delayed intervention leads to a much slower response during one or more months [Seeman, 1993]. Our model points to the possibility that maintenance therapy may have an important role not only in preventing the recurrence of positive symptoms, but that it may also slow down the progression of the disease by blocking activity-dependent changes.

Besides the intuitive notion that schizophrenic delusions and hallucinations typically arise in a spontaneous and biased manner, what other characteristics of ill-formed attractor states can be thought of as linked to the pathogenesis and manifestations of such positive symptoms? Hoffman and McGlashan provide a detailed account of the possible role of such pathological attractor states, termed 'parasitic foci', in the formation of schizophrenic positive symptoms [Hoffman & McGlashan, 1993]. Their explanations assume that parasitic foci produce their effects by altering speech perception and production processes. For example, suppose that cortical speech production regions become dominated by a parasitic focus. This may result in an experience of inner speech, which, because of the parasitic focus, is stereotyped in nature. Due to the possible detachment of such inner mental events from corresponding motor actions, these events may be experienced as unintended. This, combined with their stereotyped nature, may induce the patient to conclude that a particular alien non-self force is inserting thoughts into his head. The content of such delusions hence reflects the response of an intact rational system trying to make sense of recurrent actions occurring in the absence of an observable agent. A parasitic focus may reproduce very complex output coding for various sensory properties of an acoustic image in addition to its verbal content. In a similar fashion, a parasitic focus involving speech perception may produce a fictitious voice percept, or even mold ambiguous acoustic stimuli into its own verbal output, resulting in the production of auditory hallucinations. Along these lines, Hoffman and McGlashan describe how numerous other positive symptoms, such as ideas of reference, thought broadcasting and paranoid delusions may all be a result of parasitic foci. In a closely related spirit, Globus and Arpaia have recently proposed that due to pathological changes, the brain may settle in attractors which obtain a paranoid attunement

[Globus & Arpaia, 1994].

The occurrence of autonomous, biased memory activation in our model is the parallel of Hoffman's concept of parasitic foci. However, while our work builds upon the conceptual framework developed by [Hoffman & McGlashan, 1993], some significant points of difference should be noted. First, while Hoffman's work concentrates on investigating the effects of synaptic degenerative changes, we study the combined effects of both synaptic degeneration and regeneration, which may both have a role in the pathogenesis of schizophrenia. Second, while Hoffman's parasitic foci are mostly sub-patterns of the stored memories and may hence not be cognitively meaningful, the 'parasitic foci' in our model are the stored patterns themselves, which being cognitively meaningful are hence more likely to elucidate delusions and hallucinations. Third, while the formation of parasitic foci in Hoffman's work is coupled with memory degradation, memory is preserved in our model until late stages in the evolution of biased retrieval, due to the presence of an external input cue in the the stimulus-dependent scenario. The latter point is important since memory is generally preserved in the early stages of schizophrenia. Finally, recent cognitive studies show that the delusional and hallucinatory themes may be elucidated by a wide range of environmental cues [Hoffman & McGlashan, 1993]. This supports the notion that schizophrenic 'parasitic foci' have *a very large basin of attraction* (as the biased attractors described in this study) and are not simply fragments of independent activity (as in [Hoffman & Dobscha, 1989]).

In addition to showing that Stevens theory is plausible from a computational perspective, and that it can be realized within the framework of a neural model that accounts for some characteristics of schizophrenic positive symptoms, the current model also generates a few testable predictions:

- On the neuroanatomical level, the model can be tested by quantitatively examining the correlation between a recent history of florid psychotic symptoms and postmortem neuropathological findings of synaptic compensation in schizophrenic subjects.

- On the physiological level, the increased compensatory noise should manifest itself in increased spontaneous neural activity. While this prediction is obviously difficult to examine directly, numerous EEG studies in schizophrenics show increased sensitivity to activation procedures [Kaplan & Sadock, 1991], together with a significant increase in slow-wave delta activity which may reflect increased spontaneous activity [Jin *et al.*, 1990].

- On the clinical level, due to the formation of a large and deep basin of attraction around the memory pattern which is at the focus of spontaneous retrieval, the proposed model predicts that its retrieval (and the elucidation of the corresponding delusions or hallucinations) may be frequently triggered by various environmental cues. A recent study points in this direction [Hoffman & Rapaport, 1993]. In that study, schizophrenic patients were asked to repeat speech in which acoustic clarity was masked with superimposed multi-speaker babble. As a result, a perceptual illusion was induced in approximately 60% of the patients who reported voices, and certain words were misheard in ways that reflected the content of the hallucinated voices.

In this work we have concentrated on examining the behavior of a single cortical module. However, considering the more general scenario, where possibly many such frontal networks are involved, one still needs to explain how spontaneous memory retrieval (performed via the possible activation of many modules) remains restricted to just a few central themes, as apparent from the nature of schizophrenic delusions and hallucinations. That is, it seems that even if the retrieval of each network is concentrated on only one of its stored patterns, then, still, many such patterns may be retrieved when considering the concomitant activity of a few networks, which represent modules composing a larger frontal region. We see three possible solutions to this challenge:

1. The problem does not really exist if the whole frontal cortex is viewed as one large associative memory network. However, at least some evidence supports the notion that the frontal cortex does have a columnar-like, modular organization [Goldman & Nauta, 1977].

2. The pathological frontal synaptic changes *fully* occur in just very few frontal modules. It may well be that in most modules, the reactive synaptic changes are sufficient to restore memory retrieval capacities but do not reach the magnitude required to generate spontaneous activity (indeed, as shown in [Horn & Ruppin, 1995a], memory retrieval is already well-preserved at levels of synaptic compensation significantly lower than those required for spontaneous retrieval to emerge).

3. A pattern highly activated by the early formation of distorted, autonomous pattern retrieval (at the frontal module where it is stored) may arrest the development of a distorted retrieval distribution in other modules. We plan to investigate this hypothesis in a multi-modular model of the frontal cortex, composed of a few interacting attractor neural networks. This possibility is motivated by our finding that persistent stimulation of a

module by external cues may counteract the effect of an initial bias in memory storage and prevent the formation of a distorted distribution.

In summary, this work is another step in recent attempts to describe the workings of the brain in their most natural framework - as a neural network. In a recent commentary on [Gray *et al.*, 1991], Frith has claimed that he "would find the circuit diagrams more convincing if the verbal descriptions of how they operate were backed by a computational model" [Frith, 1991]. As this work demonstrates, a neural model may be a useful methodological tool for examining the feasibility of theoretical hypotheses within a computational context. As in previous neural models of schizophrenia, it is rather striking to realize that the disruption of just a few simple computational mechanisms can lead to rich behaviors which correspond to some of the clinical features of schizophrenic positive symptoms.

References

[1] P. Alvarez and L.R. Squire. Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl. Acad. Sci.*, 91:7041–7045, 1994.

[2] D.J. Amit. *Modeling brain function: the world of attractor neural networks*. Cambridge University Press, 1989.

[3] S. Armentrout, J. Reggia, and M. Weinrich. A neural model of cortical map reorganization following a focal lesion. *Artificial Intelligence in Medicine*, 1994. in press.

[4] P. Calabresi, R. Maj, A. Pisani, N.B Mercuri, and G. Bernardi. Long-term synaptic depression in the striatum phsiological and pharmacological charecterization. *Journal of Neuroscience*, 12 (11):4224–4233, 1992.

[5] W.T. Carpenter and R.W. Buchanan. Schizophrenia. *New England Journal of Medicine*, 330:10, 1994.

[6] S.K. Chaturvedi and V.D. Sinha. Recurrence of hallucinations in consecutive episodes of schizophrenia and affective disorder. *Schizophrenia Research*, 3:103–106, 1990.

[7] J.D. Cohen and D. Servan-Schreiber. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1):45–77, 1992.

[8] G. Dell. A spreading-activation theory of retrieval and sentence production. *Psychological Review*, 93:283–321, 1986.

[9] J.C. Eccles. The modular operation of the cerebral neocortex considered as the material basis of mental events. *Neuroscience*, 6:1839–1855, 1981.

[10] U. Frey, H. Schroeder, and H. Matthies. Dopaminergic antagonsits prevent long-term maintenance of posttetanic LTP in the CA1 region of rat hippocampal slices. *Brain Research*, 522 (1):69–75, 1990.

[11] H.R. Friedman, J.D. Janas, and P.S. Goldman-Rakic. Enhancement of metabolic activity in the diencephalon of monkeys performing working memory tasks: A 2-deoxyglucose study in behaving rhesus monkeys. *Journal of Cognitive Neuroscience*, 2(1):18–31, 1990.

[12] C. Frith. In what context is latent inhibition relevant to the symptoms of schizophrenia? *Behavioral and Brain Sciences*, 14(1):28, 1991.

[13] J.M. Fuster and J.P. Jervey. Neuronal firing in the inferotepmoral cortex of the monkey in a visual memory task. *The Journal of Neuroscience*, 2(3):361–375, 1982.

[14] G.G. Globus and J.P. Arpaia. Psychiatry and the new dynamics. *Biological Psychiatry*, 35:352–364, 1994.

[15] P.S. Goldman and W.J.H. Nauta. Columnar distribution of cortico-cortical fibers in the frontal, association, limbic and motor cortex of the developing rhesus monkey. *Brain Res.*, 122:393–413, 1977.

[16] P.S. Goldman-Rakic. Prefrontal cortical dysfunction in schizophrenia: The relevance of working memory. In B.J. Carroll and J.E. Barrett, editors, *Psychopathology and the Brain*, pages 1 – 23. Raven Press, 1991.

[17] P.S. Golman-Rakic. Cicuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum and V. Mountcastle, editors, *Handbook of Physiology*, pages 373–417. American Physiological Society, 1987.

[18] J.A. Gray, J. Feldon, J.N.P. Rawlins, D.R. Hemsley, and A.D. Smith. The neuropsychology of schizophrenia. *Behavioral and Brain Sciences*, 14:1–84, 1991.

[19] M.E. Hasselmo. Runaway synaptic modification in models of the cortex: Implications for Alzheimer's disease. *Neural Networks*, 7(1):13–40, 1994.

[20] G. Heit, M.E. Smith, and E. Halgren. Neural encoding of individual words and faces by the human hippocampus and amygdala. *Nature*, 333:773–775, 1988.

[21] M. Herrmann, E. Ruppin, and M. Usher. A neural model of the dynamic activation of memory. *Biological Cybernetics*, 68:455–463, 1993.

[22] G.E. Hinton and T. Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

[23] R. Hoffman and S. Dobscha. Cortical pruning and the development of schizophrenia: A computer model. *Schizophrenia Bulletin*, 15(3):477, 1989.

[24] R.E. Hoffman and T.H. McGlashan. Parallel distributed processing and the mergence of schizophrenic symptoms. *Schizophrenia Bulletin*, 19(1):119–140, 1993.

[25] R.E. Hoffman and J.A. Rapaport. A psycholoinguistic study of auditory/verbal hallucinations: Preliminary findings. In David A. and Cutting J., editors, *The Neuropsychology of Schizophrenia*. Erlbaum, 1993.

[26] R.E. Hoffman. Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences*, 9:503–548, 1986.

[27] R.E. Hoffman. Computer simulations of neural information processing and the schizophrenia-mania dichotomy. *Arch. Gen. Psychiatry*, 44:178, 1987.

[28] J.J. Hopfield. Neural networks and physical systems with emergent collective abilities. *Proc. Nat. Acad. Sci. USA*, 79:2554, 1982.

[29] D. Horn and E. Ruppin. Compensatory mechanisms in an attractor neural network model of schizophrenia. *Neural Computation*, 7(1):182–205, 1995.

[30] D. Horn and E. Ruppin. Synaptic maintenance in associative memory networks. 1995. Preprint.

[31] D. Horn, E. Ruppin, M. Usher, and M. Herrmann. Neural network modeling of memory deterioration in alzheimer's disease. *Neural Computation*, 5:736–749, 1993.

[32] T.M. Hyde, J.C. Ziegler, and D.R. Weinberger. Psychiatric disturbances in metachromatic leukodystrohy: Insights into the neurobiology of psychosis. *Archives of Neurology*, 49:401–406, 1992.

[33] D.C. Javitt and S.R. Zukin. Mechanisms of phenylcyclidine (pcp)-n-methyl-d-aspartate (nmda) receptor interaction: Implications for schizophrenia. In Tamminga C.A. and Schulz S.C., editors, *Advances in Neuropsychiatry and Psychopharmacology*, pages 13–19. Raven Press, 1993.

[34] Y. Jin, S.G. Potkin, D. Rice, and J. Sramek et. al. Abnormal eeg responses to photic stimulation in schizophrenic patients. *Schizophrenia Bulletin*, 16(4):627–634, 1990.

[35] H.I. Kaplan and B.J. Sadock. *Synopsis of Psychiatry.* Williams and Wilkins, 1991.

[36] R. Karler, L.D. Calder, and S.A. Turkanis. Dnqx blockade of amphetamine behavioral sensitization. *Brain Research*, 552:295–300, 1991.

[37] M. Marsel Mesulam. Schizophrenia and the brain. *New England Journal of Medicine*, 322(12):842–845, 1990.

[38] Y. Miyashita and H.S. Chang. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, 331:68–71, 1988.

[39] P. Monowitz, A. Kling, and H. Kohn. Clinical course of adult metachromatic leukodystrophy presenting as schizophrenia: A report of two living cases in siblings. *The Journal of Nervous and Mental Disease*, 166(7):500–506, 1978.

[40] V.B. Mountcastle. An organizing principle for cerebral function: The unit module and the distributed system. In F.O. Schmitt and F.G. Worden, editors, *The Neurosciences: Fourth Study Program*, pages 21–42. MIT Press, 1979.

[41] W.J.H. Nauta and V.B. Domesick. Neural associations of the limbic system. *The neural basis of behavior*, 10:175–206, 1982.

[42] J.W. Petegrew, M.S. Keshavan, K. Panchalingam, S. Strychor, D.B. Kaplan, M.G. Tretta, and M. Allan. Alterations in brain high-energy phosphate and membrane phospholipid metabolism in first-episode, drug-naive schizophrenics: A pilot study of the dorsal prefrontal cortex by in vivo phosphorous 31 nuclear magnetic spectroscopy. *Arch. of General Psychiatry*, 48:563–568, 1991.

[43] J.W. Petegrew, M.S. Keshavan, and N.J. Minshew. Nuclear magnetic resonance spectroscopy: neurodevelopment and schizophrenia. *Schizophrenia Bulletin*, 19:35–53, 1993.

[44] J.A. Reggia, P. Marsland, and R.S. Berndt. Competitive dynamics in a dual-route connectionist model of print-to-sound transformation. *Complex Systems*, 2:509–547, 1988.

[45] J. Reggia, R. Berndt, and L. D'Autrechy. Connectionist models in neuropsychology. In *Handbook of Neuropsychology*, volume 9. 1994.

[46] E. Ruppin and J. Reggia. A neural model of memory impairment in diffuse cerebral atrophy. *Br. Jour. of Psychiatry*, 166(1):19–28, 1995.

[47] E. Ruppin. Neural modeling of psychiatric disorders. *Network: Computation in Neural Systems*, 1995. Invited review paper, to appear.

[48] K. Sakay and Y. Miyashita. Neural organization of the long term memory of pair associates. *Nature*, 354:152–159, 1990.

[49] P. Seeman. Schizophrenia as a brain disease: The dopamine receptor story. *Arch. Neurol.*, 50:1093–1095, 1993.

[50] D. Servan-Schrieber, H. Printz, and J.D. Cohen. A network model of cathecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, 249:892–895, 1990.

[51] A.D. Sherman, A.T. Davidson, S. Baruah, T.S. Hedgewood, and R. Waziri. Evidence of glutamatergic deficiency in schizophrenia. *Neuroscience Letters*, 121:77–90, 1991.

[52] L. R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99:195–231, 1992.

[53] J.R. Stevens. Abnormal reinnervation as a basis for schizophrenia: A hypothesis. *Arch. Gen. Psychiatry*, 49:238–243, 1992.

[54] M.V. Tsodyks and M.V. Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.*, 6:101 – 105, 1988.

[55] D.R. Weinberger. Anteromedial temporal-prefrontal connectivity: A functional neuroanatomical system implicated in schizophrenia. In B.J. Carroll and J.E. Barrett, editors, *Psychopathology and the Brain*, pages 25–43. Raven Press, 1991.

[56] P.C. Williamson. Schizophrenia as a brain disease. *Arch. Neurol.*, 50:1096–1097, 1993.

[57] F.A.W. Wilson, S.P.O Scalaidhe, and P.S. Goldman-Rakic. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, 260:1955–1958, 1993.

Appendix A: **A Formal Description of the Model**

We use a biologically-motivated variant of Hopfield's attractor neural network model, proposed by Tsodyks and Feigelman [1988]. The network is composed of N neurons, where each neuron i is described by a binary variable $S_i = \{1, 0\}$ denoting an active (firing) or passive (quiescent) state, respectively. All neurons have a fixed, uniform, optimally-tuned threshold θ [Horn & Ruppin, 1995a]. M distributed memory patterns ξ^μ, where superscript μ indicates a pattern index, are stored in the network. The elements of each memory pattern are randomly chosen to be 1 (0) with probability p $(1-p)$, respectively, with $p \ll 1$.

In each set of parameters characterizing a given network, the behavior of the network is monitored over many trials. In each trial, the initial state of the network $S(0)$ is random, with average activity level $q < p$, reflecting the notion that the network's baseline level of activity is lower than its activity in the persistent memory states. The neuron's state is updated stochastically, in accordance with its input. The input (post-synaptic potential) h_i of neuron i at time t is the sum of internal contributions from other neurons in the network and external contribution ${F_i}^e$, given by

$$h_i(t) = \sum_j W_{ij} S_j(t-1) + {F_i}^e \ . \tag{1}$$

The updating rule for neuron i at time t is given by

$$S_i(t) = \begin{cases} 1, & \text{with probability } G(h_i(t) - \theta) \\ 0, & \text{otherwise} \end{cases} \quad , \tag{2}$$

where G is the sigmoid function $G(x) = 1/(1 + \exp(-x/T))$, T denotes the noise level, and θ is the neuron's threshold, which is optimally tuned to guarantee perfect retrieval in the network's premorbid state [Horn & Ruppin, 1995a].

For each pattern ξ^μ that is stored in the network, the synaptic connections are modified in accordance with the Hebbian rule

$$W_{ij}^{new} = W_{ij}^{old} + \frac{c}{N}({\xi^\mu}_i - p)({\xi^\mu}_j - p) \ , \ i, j = 1 \cdots N \ , \ j \neq i \ . \tag{3}$$

Before any pattern is stored, the synaptic matrix weights are taken to be zero. The values of the parameters c, T and e (see below) determine the synaptic strengths in the network, as synaptic deletion ($e < e_0$) and compensation ($c > c_0$) take place.

The behavior of the network is examined in two scenarios. In the *stimulus-dependent retrieval* scenario, a stored memory pattern (say ξ^1) is presented as an input cue to the network via the external synaptic projections, such that

$$F_i^e = e \cdot \xi^1{}_i \quad (e > 0) \ . \tag{4}$$

Following the dynamics defined in (1) and (2), the network state evolves until it converges to a stable state. Performance is then measured by the similarity between the network's end state S and the cued memory pattern ξ^μ (which is the desired response), conventionally denoted as the *overlap* m^μ, and defined by

$$m^\mu = \frac{1}{p(1-p)N} \sum_{i=1}^{N} (\xi_i^\mu - p) S_i \ . \tag{5}$$

In addition to investigating the network's activity in response to the presentation of an external input, we also examine its behavior in the absence of any specific stimulus. In this case, the network may either continue to wander around in a state of random low baseline activity, or it may converge onto a stored memory state. We refer to the latter process as *spontaneous retrieval.*

The hypothesized activity-dependent schizophrenic pathological synaptic changes are modeled via the rule

$$W_{ij}(t) = W_{ij}(t-1) + \frac{\gamma}{N} (\bar{S}_i - p)(\bar{S}_j - p) \ , \tag{6}$$

where t is a time index (i.e., number of iterations the network undergoes), $\bar{S}_k$ is 1 (0) only if neuron k has been consecutively firing (quiescent) for the last τ iterations, and γ is a constant determining the magnitude of activity-dependent changes. If either of the neurons i or j has not remained in the same firing state in all of the last τ iterations, then the synaptic weight W_{ij} is not modified. This activity-dependent synaptic modification mechanism is a much simplified model of long-term potentiation and long-term depression processes, where a train of impulses is required for a synaptic modification to occur.

It should be noted that the activity-dependent synaptic modification mechanism (6) reduces to the learning algorithm which generates the synaptic matrix (3) when the external inputs are the memorized patterns and the network is in its intact, premorbid state. One only has to make the additional requirement that the external projections' strength during the learning stage is sufficient to align the network's activity with the pattern to be stored.

Appendix B: **Explaining the Observed Phenomena**

We now present some computational considerations which explain why Hebbian-like synaptic changes lead to the generation of a biased retrieval distribution in the spontaneous retrieval mode, and why an end-state mixed attractor is eventually formed. We also define the (quite broad) class of attractor neural models for which our results are valid.

The concentration of spontaneous retrieval on one memory pattern is an expression of a property of the network known in physics as 'spontaneous symmetry breaking': as some memory pattern is spontaneously retrieved, its corresponding 'basin of attraction' is further enlarged due to Hebbian activity-dependent modification of the synaptic matrix. This follows since the 'energy' level $-\sum_{ij} S_i S_j W_{ij}$ of a spontaneously retrieved memory pattern strictly decreases after the internal synapses have been modified via expression (6), and the probability of convergence to an attractor from a random initial state increases exponentially with the absolute magnitude of its 'energy' level. Via this exponential positive feed-back loop, any initial bias in the network's initial state would break the symmetry underlying the original synaptic memory matrix, and lead to an inhomogeneous distribution of spontaneously retrieved states. This is not the case when external inputs drive the network, and the expression of the spontaneous feedback loop is suppressed.

To understand how the global mixed-attractor is eventually formed, one should first note that the non-Hebbian synaptic compensatory changes, which initially lead to the generation of spontaneous retrieval, have an overall excitatory effect, as shown in [Horn & Ruppin, 1995a]. As a result of this excitatory effect, the average fraction of firing neurons r in the stable states which the network converges to during spontaneous retrieval, is larger than the original coding fraction p. When activity-dependent changes occur concomittantly with the generation of spontaneous retrieval, the expected value of synaptic modification performed via expression (6) is

$$E(\Delta(W_{ij})) = r^2(1-p)^2 + (1-r)^2p^2 - 2r(1-r)p(1-p) = (r-p)^2 > 0 \; . \qquad (7)$$

The net change in the network is always excitatory, and this consideration seems to lead to the conclusion that the network should always end up eventually in a global excitatory state in which all neurons fire, but this is not the case. Due to the positive feedback loop described above, the network may repeatedly converge to the same stable state in a series of consecutive trials. When this occurs, we may calculate separately the expected value of synaptic modification of the synapses connected to a (currently) firing neuron, and that of those connected to a quiescent one; for a firing neuron we obtain

$$E(\Delta(W_{ij})) = (r - p)(1 - p) > 0 \ , \tag{8}$$

and for a quiescent neuron

$$E(\Delta(W_{ij})) = p(p - r) < 0 \ . \tag{9}$$

Hence, the average input field of the firing neurons tends to become more positive and that of the quiescent ones more negative. Although the synaptic changes have a global excitatory effect, their distinct effect on the firing and quiescent neurons tends to further stabilize them. This description, however, hinges upon the assumption that the final stable states generated in consecutive trials are similar. Due to thermal noise and the random initial baseline activity, the network may converge at time to a different stable state, and break the streak of self-stabilizing, activity-dependent changes. When this occurs, the synaptic excitatory modifications, following (7), result in mixed states with increasing activity levels r. Note that, as times evolves, the synaptic magnitudes are increased and the effects of the thermal fluctuations vanish. Hence, the later in the evolution of spontaneous retrieval a stable state appears, the more likely it is to remain stable; this explains the observed considerable stability of the 'end-state' mixed attractors. Similar arguments provide an intuitive account for the network's behavior when synaptic bounds are enforced. The bounds are sufficiently large so that during the evolution of spontaneous retrieval the network arrives at stable states that are stable enough to be repeated. It then follows from (6) that the synaptic modifications tend to increase the magnitude of both excitatory and inhibitory weights, and keep the network's weights in the vicinity of the bounds.

The arguments above pertain to the class of attractor neural network models which have the property that uniform synaptic compensation has an excitatory effect on the network activity. This condition ensures that spontaneous activity will emerge (as shown in [Horn & Ruppin, 1995a]), and that $r > p$. As shown now, in this state, once spontaneous retrieval emerges and Hebbian changes are incorporated, then retrieval will eventually concentrate on a single biased memory, until a global mixed attractor is formed. The notion that compensatory synaptic strengthening has an overall excitatory effect seems rather plausible from a biological point of view, and hence this class of models is of interest.

In the simple model presented in this paper, each trial ends after the network has converged to a stable state, or after it has remained wandering around in its baseline low-activity state for some time. A more realistic scenario should include some mechanism that enables the network to revert from its attractor states back to its baseline random state, and then another trial may begin (in the current framework we have reset the network's state at the beginning of every new trial). Disregarding the precise mechanism which is actually utilized (see, for example, [Herrmann *et al.*, 1993]), it is a known general property of attractor neural networks that as the energy level of a state becomes lower its stability increases. Hence, once attracted into a state towards which retrieval is biased, the network will tend to remain in that state for a much longer period than when it converges to an 'unbiased' memory state.