Pathogenesis of Schizophrenic Delusions and Hallucinations: A Neural Model

by Eytan Ruppin, James A. Reggia, and David Horn

Abstract

We implement and study a computational model of Stevens' theory of the pathogenesis of schizophrenia. This theory hypothesizes that the onset of schizophrenia is associated with reactive synaptic regeneration in brain regions that receive degenerating temporal lobe projections. Focusing 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 similar to 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 delusions and hallucinations that occur in schizophrenia without any apparent external trigger and for their tendency to concentrate on a few central cognitive and perceptual themes. The model explains why the positive symptoms of schizophrenia tend to wane as the disease progresses, why delayed therapeutic intervention leads to a much slower response, and why delusions and hallucinations may persist for a long time when they do occur.


Neural modeling research is currently a 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; Hasselmo 1994; Ruppin and Reggia 1995), and cognitive and behavioral explorations of aphasia and acquired dyslexia (Dell 1986; Reggia et al. 1988; Hinton and Shallice 1991).

Some pioneering steps in modeling schizophrenia have also been undertaken recently. The two main directions have been the modeling of positive symptoms in schizophrenia and the modeling of various aspects of cognitive functions in persons with schizophrenia. The first avenue, taken by Hoffman, concentrated on modeling positive symptoms in the framework of an associative memory network (Hoffman 1987; Hoffman and Dobscha 1989). This work pointed to a possible link between the appearance of specific neurodegenerative changes and the emergence of...
parasitic foci,” states in which a neural network’s normal processing is disrupted and locked into dysfunctional patterns of activity. Clinical experiments then explored the possible role of such parasitic foci in the formation of positive symptoms (Hoffman and McGlashan 1993). Working in a second modeling framework, Cohen and Servan-Schreiber provided a detailed computational account explaining how some functional deficits can arise from the neuro-modulatory effects of dopamine hypothesized in schizophrenia (Servan-Schreiber et al. 1990; Cohen and Servan-Schreiber 1992).

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 neuro-anatomical and cognitive terms, may be realized within the framework of a detailed (albeit simplified) neural model, and if so, what possible insights 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 on the pathophysiology of schizophrenia suggests that there are atrophic changes in the hippocampus and parahippocampal areas in the brains of a significant number of patients with schizophrenia, 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 led Stevens to hypothesize that the onset of schizophrenia is associated with reactive anomalous sprouting and synaptic reorganization in the projection sites of dystrophic medial temporal neurons.

To study the possible functional implications of Stevens’ hypothesis, we studied 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 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 as the focus of our study primarily because a considerable amount of data shows that there is frontal lobe involvement in schizophrenia (see Weinberger [1991] for a comprehensive review). Moreover, these data suggest that prefrontal hypometabolism in schizophrenia may be secondary to the function of aberrant afferent temporal projections (Weinberger 1991): Recent in vivo metabolic data indicate that pre-frontal cortex and hippocampus are functionally coupled during memory tasks (Friedman et al. 1990) and that sharing information via numerous temporal-frontal connections is an important aspect of prefrontal functioning (Goldman-Rakic 1987). Overall, it seems that the disruption of temporofrontal connections and reactive frontal sprouting may have an important functional role in the pathogenesis of schizophrenia. From an anatomical perspective, there are both direct routes to frontal lobes from medial temporal lobe structures and indirect routes via the thalamus and nucleus accumbens (Nauta and Domesick 1982). For simplicity, we are focusing on the possible functional effects of damage in the direct route, which may also approximate the cascading effect of weakening temporosubcortical connections on indirect temporofrontal routes.

The frontal cortex is a plausible site of working memory in the brain (Goldman-Rakic 1991) and, together with other areas of associative cortex, may also be an important storage site for long-term associative, content-addressable memory. We therefore assume that memory retrieval from the frontal cortex is invoked by the firing of incoming temporal projections. This assumption is based on 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 and Squire 1994) and on 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
Figure 1. Schematic illustration of the model

Each frontal module is modeled as an attractor neural network whose neurons receive inputs via 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.

Patterns in a Hebbian, internal, synaptic matrix. Such a frontal module represents a macrocortical unit, which has been suggested as a basic functional building block of the neocortex (Goldman and Nauta 1977; Mountcastle 1979; Eccles 1981). This unit receives specific memory-cue inputs from temporal projections and is connected with other areas via diffuse intermodular connections. In accordance with Stevens’ theory, when pathological processes occur in schizophrenia the degeneration of temporal projections is modeled by decreasing the strength of the incoming temporofrontal 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 affecting the network dynamics. The latter is obviously a crude simplification of the possible role of the intercortical projections. But it is partially justified since cue input patterns are assumed to arise specifically from temporal lobe projections, so 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.

Methods

Our model views 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 (Hopfield 1982; Amit 1989), 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, that is, made to be “attractors” of the network dynamics. In other words, if a pattern that 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 reaches a state representing that memory pattern. Such a network could be regarded as an associative memory system.

In attractor neural networks, stored memories are not represented at specific neurons of the network, but their corresponding representations are distributed; many neurons participate in a given 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 and Jervey 1982; Miyashita and Chang 1988) and frontal (Wilson et al. 1993) cortices. These experiments show that some cortical neurons continue to fire at an increased rate for a few seconds even after the external stimulus that originally triggered the 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 reflect a collective behavior (Miyashita and Chang 1988; Sakay and Miyashita 1990).

We use a biologically motivated variant of an attractor neural network model, proposed by Tsodyks and Feigelman (1988). The network is composed of N neurons, where each neuron may be in either an active (firing) or passive (quiescent) state. Each neuron's state is updated stochastically, in accordance with its synaptic inputs (membrane potential), that is, the signals it receives from the active neurons in the network and from external, intermodular sources. If a neuron's current membrane potential is significantly higher than its firing threshold, that neuron's state will be active; otherwise it is likely to remain silent (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 on the sign and magnitude of these synapses.

As illustrated in figure 1, the neurons receive three kinds of connections. (1) **External input connections** represent 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 c. (2) **Internal connections** that store M memorized patterns and represent the intramodular frontal connections. The magnitude of these connections is 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 the strength parameter c of the internal connections. (3) **External diffuse connections** represent "nonspecific" intermodular connections. The latter's effect on network dynamics is expressed by 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. In the **stimulus-dependent retrieval** scenario, a stored memory pattern is presented as an input cue to the network via the external synaptic projections, and 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 become similar to any of the stored memory patterns.

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, a process known as **spontaneous retrieval**. In this scenario, the external input synapses are inactive 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 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 into a stored memory state, resulting in a pathological, autonomous activation of memorized patterns.

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. A memory pattern is considered to be retrieved if the network converges to a stable state with which it has similarity greater than 0.9. 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 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 \( \epsilon = 0.035 \), 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, that is, the retrieval performance of the network is near perfect.

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' (1992) theory, the external input synapses' magnitude (\( \epsilon \)) is weakened and the internal synapses magnitude (\( c \)) and noise level (\( T \)) are increased.

First, in the experiments described in the following section, Emergence of Spontaneous Retrieval, 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, increasing the internal synaptic strength or the noise level serves 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 the Biased Spontaneous Retrieval section, we assume 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 by which the memorized patterns were initially stored. Thus, we assume that the primary pathological process proposed by Stevens—the degeneration of external synapses—triggers a compensatory increase in internal memory-bearing synapses, and that the latter has combined nonactivity-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 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 delusions and hallucinations in schizophrenia.

Emergence of Spontaneous Retrieval. After weakening the external synapses to the level \( \epsilon = 0.015 \), 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 noncued memory patterns. Beyond some critical level of increase in either the internal synaptic strength or the noise level, the network begins to retrieve memory patterns frequently even though it does not receive any external input!

Schizophrenic symptomatology involves complicated cognitive and perceptual phenomena that require much more elaborate representations than a simple neural model of associative memory. Yet delusions and hallucinations in schizophrenia frequently appear in the absence of any apparent external trigger. It therefore seems plausible that the spontaneous activation of stored patterns described above is an essential element in their pathogenesis. It should also be noted that when spontaneous retrieval emerges, the network may spontaneously converge at times to nonstored patterns, which are a
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 ($\epsilon = 0.015$) 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 $c$; $T = 0.009$

mixture of a few memory patterns (Horn and Ruppin 1995). Retrieval of mixed patterns may help explain the generation of more complex forms of delusions and hallucinations.

Using Stevens' (1992) hypothesis to model the regenerative synaptic changes, what then is the computational role of increased internal synaptic strength and noise level? 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 state 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 100 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 $\epsilon$. For fixed $T$, performance decreases rapidly as the external input strength $\epsilon$ is decreased. However, if the decrease in $\epsilon$ 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 retrieval performance at some $\epsilon$-
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 input strength. As $c$ 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 already achieved 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 the face of the weakened external input synapses (representing degenerated temporal projections).

**Biased Spontaneous Retrieval.** We now turn to the effects of incorporating Hebbian, activity-dependent synaptic changes into internal synaptic regeneration. This investigation is motivated by recent findings that increased dopaminergic activity may enhance Hebbian-like activity-dependent synaptic changes (see Discussion section) and that the density of N-methyl-D-aspartate (NMDA) receptors is increased in cortical areas of patients with schizophrenia (Javitt and Zukin 1993). It should be emphasized that the activity-dependent synaptic modification mechanism we use is essentially the same Hebbian activity-dependent mechanism used to store memorized patterns dur-
ing a learning episode (appendix A). We thus assume that the synaptic modification rate, denoted $\gamma$, has significant magnitude during the early “childhood” plastic period. It later decreases to near zero, a level that is 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 used 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 from a different initial random state. Because of the compensatory synaptic changes, some of the memorized patterns are now spontaneously retrieved, and because of the incorporation of activity-dependent synaptic changes ($\gamma = 0.0025$), the synaptic matrix is no longer fixed. 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, affects the future dynamic behavior of the network.

Figure 4 traces the distribution of the memory patterns the network has spontaneously converged to after the first 100 trials preceding the 200th trial, the 500th trial, and the 800th trial. The total frequency of convergence to memory patterns increases with the number of 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 spontaneously retrieved memory patterns tends to concentrate on a single memory pattern as more trials occur. Although the synaptic matrix was initially nonbiased, small, random correlations between the net-
The biased spontaneous retrieval of memory patterns is sufficient to overwhelmingly enhance their retrieval. Thus, biased retrieval develops, and out of the many patterns stored in the network only a few are spontaneously retrieved. The biased spontaneous retrieval distribution that develops 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 on a limited set of recurring cognitive and perceptual themes (e.g., Hoffman 1986; Chaturvedi and Sinha 1990).

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 it does not have high overlap with any memorized pattern and is thus considered not to represent 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 a 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 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 than 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 toward the biased memory (#1). Thus, when spontaneous retrieval emerges, the network has a strong tendency to markedly amplify preexisting biases in its internal synaptic matrix.

Next we examine the distribution of memories retrieved when the network operates in the stimulus-dependent retrieval scenario. The network is similar to the one used above, but now a memory pattern is randomly chosen and presented as an external input cue to the network on each trial. 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 loses 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 up to levels of initial bias of $k = 2.5$. Thus, the external input, which is homogeneously distributed among the memorized patterns, counteracts 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 later stages in the evolution of the synaptic matrix, when a mixed attractor is formed. Hence, while synaptic regenerative changes may lead to 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 strongly biased, that is, 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 $k = 3.5$. Thus, it appears that the mechanisms leading to the formation of biased spontaneous retrieval are geared primarily toward 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 do not have unlimited effectiveness, obviously raises the question of the network's behavior in the presence of a limit on the absolute synaptic magnitude. Introducing a limit on the absolute magnitude of the weights of the internal synaptic matrix (of value 2.5), we have rerun the simulations described...
Figure 5. The distribution of stimulus-dependent retrieval of memories

Discussion

We have shown that while preserving memory performance, compensatory synaptic regenerative changes modeling those proposed in Stevens' (1992) 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 non-homogeneous retrieval distribution that is strongly dominated by a single memory pattern. A small initial bias in the memory matrix toward 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 intramodular synaptic connections. Both types 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 our results testify to the plausibility of Stevens' theory by showing that the realization of the hypothesized pathological synaptic changes within a neural...
model leads to significant changes in the macroscopic behavior of the network, which share some of the characteristics of positive symptoms of schizophrenia.

It should be emphasized that spontaneous, biased retrieval in the model cannot emerge 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 together within a certain range of synaptic strengths, that autonomous pathological activation of stored patterns emerges. In contrast to what one may intuitively think, not many changes in the network parameters can produce these pathological manifestations and also preserve memory retrieval. To begin with, analytic considerations show that the emergence of spontaneous retrieval requires considerable strengthening of internal or diffuse external synapses (Horn and Ruppín 1995). In addition, as demonstrated in this article, a spontaneous retrieval distribution will become biased only if the internal synaptic strengthening includes an activity-dependent component and if the frequency of successful stimulus-dependent retrieval is markedly reduced. Moreover, 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 $\gamma$ 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 pathological aberrations. 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 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 raises the question of how the brain normally resists generating such pathological attractors in its healthy state. We believe that several defensive mechanisms 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 nonbiased) 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. Thus, 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 and Ruppín, in preparation).

Our results provide considerable support for 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 about the pathological changes underlying schizophrenic symptomatology. On one hand, a broad range of findings supports Stevens’ proposal (1992) that synaptic regeneration occurs in target areas of degenerating temporal projections including, for example, increased glutamate uptake sites, expansion of NMDA binding sites, increased dendritic branching of pyramidal cells, and increased levels of synaptophysin-like 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 patients with schizophrenia (Pettegrew et al. 1991, 1993) and that synaptic glutamate is reduced in prefrontal regions in schizophrenia (Sherman et al. 1991). While on a large scale there is an overall reduction of synapses, it is possible that local "islands" of excess synaptic regeneration can result in spontaneous...
retrieval in a few cortical modules. A metachromatic leukodystrophy model of schizophrenia (Hyde et al. 1992) lends support to Stevens' idea that a disconnection between temporal and frontal systems plays an important role in the pathogenesis of schizophrenia. In a number of cases with a rare adult type of metachromatic leukodystrophy, the disease initially presents with schizophrenia-like psychosis. These patients may be psychiatrically treated for years before neurological symptoms become manifest (Monowitz et al. 1978). As noted by Weinberger (1991) and Hyde and colleagues (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 may be that, in the absence of significant neuronal damage, psychotic symptoms arise because of synaptic disconnection.

Despite a number of suggestive findings, there is currently no proof that a global abnormality of neurotransmission is a primary feature of schizophrenia (Mesulam 1990; Williamson 1993; Carpenter and 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-Schreiber et al. 1990; Cohen and 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 reducing the positive symptoms of schizophrenia 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 subserving 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 (Calaresi 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 positive symptoms of schizophrenia. First, the emergence of spontaneous, nonhomogeneous retrieval is a self-limiting phenomenon; eventually, a global mixed-state attractor is formed that is not similar to 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 (Hoffman and McGlashan 1993; Globus and Arpaia 1994). This parallels the clinical observation that as schizophrenia progresses, positive symptoms tend to wane while negative symptoms are enhanced (Gray et al. 1991; Kaplan and Sadock 1991; Carpenter and 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 phase, when spontaneous retrieval emerges and activity-dependent changes have not yet made their mark (see Emergence of Spontaneous Retrieval section). While these 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 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 nature of positive symptoms, which may endure for long periods. Third, as more and more spontaneous retrieval trials occur, the frequency of spontaneous re-
traversal increases until some point is reached where it declines sharply. A similar pattern should be observed in the frequency of positive symptoms in schizophrenia 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, delayed intervention leads to a much slower response during 1 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 the progression of the disease by blocking activity-dependent changes.

In addition to the intuitive notion that the delusions and hallucinations of schizophrenia typically arise in a spontaneous and biased manner, what other characteristics of ill-formed attractor states might be linked to the pathogenesis and manifestations of such positive symptoms? Hoffman and McGlashan (1993) provide a detailed account of the possible role of such pathological attractor states, or parasitic foci, in the formation of positive symptoms. 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. An experience of inner speech may result, which, because of the possible detachment of such inner mental events from corresponding motor actions, these events may be experienced as unintended. Thus, the patient may conclude that an alien nongenome force is inserting thoughts into his or her head. The content of such delusions reflects the response of an intact rational system trying to make sense of recurrent actions occurring in the absence of an observable agent. In addition to its verbal content, a parasitic focus may reproduce complex output coding for various sensory properties of an acoustic image. 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 (1993) 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 vein, Globus and Arpaia (1994) recently proposed that pathological changes may cause the brain to settle in attractors that obtain a paranoid attunement.

The occurrence of autonomous, biased memory activation in our model is parallel to Hoffman's concept of parasitic foci. However, while our work builds on the conceptual framework developed by Hoffman and 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 synaptic degeneration and regeneration, both of which may play a role in the pathogenesis of schizophrenia. Second, Hoffman's parasitic foci are mostly subpatterns of the stored memories and hence may not be cognitively meaningful. But the parasitic foci in our model are the stored patterns themselves, which, being cognitively meaningful, are more likely to elucidate delusions and hallucinations. Third, while the formation of parasitic foci in Hoffman's work is coupled with memory degradation, memory in our model is preserved until late in the evolution of biased retrieval because of the presence of an external input cue in the stimulus-dependent scenario. The latter point is important because 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 and McGlashan 1993). This supports the notion that parasitic foci in schizophrenia have a very large basin of attraction, like the biased attractors described in this study, and are not simply fragments of independent activity, as in Hoffman and Dobscha (1989).

In addition to showing that Stevens' theory (1992), 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 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 subjects with schizophrenia.
- 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 electroencephalographic studies in patients with schizophrenia show increased sensitivity to activation procedures (Kaplan and Sadock 1991) and 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 that is the focus of spontaneous retrieval, the proposed model predicts that its retrieval (and the elucidation of the corresponding delusions or hallucinations) frequently may be triggered by various environmental cues. A recent study by Hoffman and Rapaport (1994) points in this direction. In that study, patients with schizophrenia were asked to repeat speech in which acoustic clarity was masked with superimposed multispeaker babble. A perceptual illusion was induced in approximately 60 percent 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 concentrated on examining the behavior of a single cortical module. However, considering the more general scenario, where many such frontal networks may be 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 is 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, many such patterns may still 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 and Nauta 1977).

2. The pathological frontal synaptic changes fully occur in just a 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 and Ruppin (1995), 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 multimodal model of the frontal cortex composed of a few interacting attractor neural networks. This examination 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 the work of Gray et al. (1991), Frith (1991) claimed that he "would find the circuit diagrams more convincing if the verbal descriptions of how they operate were backed by a computational model" (p. 28). As this work demonstrates, a neural model may be a useful tool for examining the feasibility of theoretical hypotheses within a computational context. As in previous neural models of schizophrenia, it is striking that the disruption of just a few simple computational mechanisms can lead to rich behaviors that correspond to some of the clinical features of schizophrenia's positive symptoms.

References


Calabresi, P.; Maj, R.; Pisani, A.; Mercuri, N.B.; and Bernardi, G. Long-term synaptic depression in the striatum: Physiological and


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


Horn, D., and Ruppin, E. "Synaptic Maintenance in Associative Memory Networks." In preparation.


Weinberger, D.R. Anteromedial temporal-prefrontal connectivity: A functional neuroanatomical system implicated in schizophrenia. In:
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 \( \Theta \) (Horn and Ruppin 1995). \( M \) distributed memory patterns \( \xi^\mu \), where superscript \( \mu \) 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 \( p < 1 \).

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 \( a < p \), 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 (postsynaptic potential) \( h_i \) of neuron \( i \) at time \( t \) is the sum of internal contributions from other neurons in the network and external contribution \( F_i \), given by

\[
h_i(t) = \sum_{\ell} W_{i\ell} S_{\ell}(t - 1) + F_i \quad (1)
\]

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

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

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

For each pattern \( \xi^\mu \) that is stored in the network, the synaptic connections are modified in accordance with the Hebbian rule

\[
W_{\xi^\mu i} = W_{\xi^\mu i} + \frac{c}{N} (\xi_i^\mu - p) \\
(\xi_j^\mu - p) i,j = 1 \ldots N, j \neq i \quad (3)
\]

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, $\xi'$) is presented as an input cue to the network via the external synaptic projections, such that

$$F_i = e \cdot \xi_i', \quad (e > 0) \quad (4)$$

Following the dynamics defined in equations (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 $\xi'$ (which is the desired response), conventionally denoted as the overlap $m^\nu$, and defined by

$$m^\nu = \frac{1}{p(1 - p)N} \sum_{i=1}^{N} (\xi'^i - p)S_i \quad (5)$$

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

$$W_{ji}(t) = W_{ji}(t - 1) + \gamma \frac{\tau}{N} (S_j - p)(S_i - p) \quad (6)$$

where $t$ is a time index (i.e., number of iterations the network undergoes), $S_k$ is 1 (0) only if neuron $K$ has been consecutively firing (quiescent) for the last $\tau$ iterations, and $\gamma$ 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 $\tau$ iterations, then the synaptic weight $W_{ji}$ 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, equation (6), reduces to the learning algorithm that generates the synaptic matrix, equation (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 to 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 $-\sum S_i S_j W_{ij}$ of a spontaneously retrieved memory pattern strictly decreases after the internal synapses have been modified via equation (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 feedback loop, any initial bias in the network’s state would break the symmetry, underlying the original synaptic memory ma-
network and the expression of the when external inputs drive the excitatory changes, which initially lead the non-Hebbian synaptic compensatory changes, which initially lead to the generation of spontaneous retrieval, have an overall excitatory effect, as shown in Horn and Ruppin (1995). As a result of this excitatory effect, the average fraction of firing neurons $r$ in the stable states that the network converges to during spontaneous retrieval is larger than the original coding fraction $p$. When activity-dependent changes occur concomitantly with the generation of spontaneous retrieval, the expected value of synaptic modification performed via equation (6) is

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

and for a quiescent neuron

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

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 on the assumption that the final stable states generated in consecutive trials are similar. Thermal noise and the random initial baseline activity may cause the network to converge at times to a different stable state, thus breaking the streak of self-stabilizing, activity-dependent changes. When this occurs, the synaptic excitatory modifications, following equation (7), result in mixed states with increasing activity levels $r$. Note that, as time 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 explanation for the network's behavior when synaptic limits are enforced. The limits are sufficiently large so that during the evolution of spontaneous retrieval the network arrives at stable states stable enough to be repeated. It then follows from equation (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 limits.

The arguments above pertain to the class of attractor neural network models in which uniform synaptic compensation has an excitatory effect on the network activity. This condition ensures that spontaneous activity will emerge (see Horn and Ruppin 1995) and that $r > p$. In this state, once spontaneous retrieval emerges and Hebbian changes are incorporated, 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 models presented in this article, 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 so that another trial may begin. (In the current framework we have reset the network's state at the beginning of every new trial.) Regardless of the precise mechanism used (e.g., Herrmann et al. 1993), it is a known general property of attractor neural networks that as the energy level of a state decreases, its stability increases. Hence, once it is attracted into a state toward 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.