Neural Modeling of Psychiatric Disorders

Eytan Ruppin *
Department of Computer Science
University of Maryland, College Park, MD 20742
ruppin@cs.umd.edu

Departments of Computer Science and Physiology
Tel-Aviv University, Tel-Aviv, Israel, 69978
ruppin@math.tau.ac.il

March 25, 1995

Abstract

This paper reviews recent neural modeling studies of psychiatric disorders. Numerous aspects of psychiatric disturbances have been investigated, such as the role of synaptic changes in the pathogenesis of Alzheimer's disease, the study of spurious attractors as possible neural correlates of schizophrenic positive symptoms, and the exploration of the ability of feed-forward and recurrent networks to quantitatively model the cognitive performance of schizophrenic patients. Current models all employ considerable simplifications, both on the level of the behavioral phenomenology they seek to explore, and on the level of their structure and dynamics. However, it is encouraging to realize that the disruption of just a few simple computational mechanisms can lead to behaviors which correspond to some of the clinical features of psychiatric disorders, and can shed light on their pathogenesis.

1 Introduction

Neural modeling research is currently a very active and growing scientific field with intense, multidisciplinary activity. Building upon the investigation of cognitive and neural functions in normal, healthy subjects, there has recently been a growing interest in the use of neural models to investigate brain pathologies and their cognitive and behavioral effects. For example, starting from 1988, the National Institutes of Mental Health in the

*The preparation of this review was supported in part by a Rothschild Fellowship.

United States has initiated an ongoing research program in Computational, Theoretical and Mathematical Neuroscience, and the first international workshop on Neural Modeling of Cognitive and Brain disorders, sponsored by the National Institutes of Health, is to be held soon (Maryland, June 1995). The increased interest of the psychiatric and psychological communities in neural network modeling (see, e.g., [Cohen & Servan-Schreiber, 1992b; Mandell & Selz, 1992; Globus & Arpia, 1994; Callaway *et al.*, 1994] reflects at least partially the feeling that even though the biological approach based on gathering basic neurobiological data has led to much progress in our understanding of basic brain mechanisms, we do not appear to have come much closer to understand how these mechanisms result in behavior [Cohen & Servan-Schreiber, 1992b]. As succinctly stated by [Mandell & Selz, 1992], we now know quite a lot about 'what' but not much about 'how'. The goal of this paper is to bring the current research done on modeling psychiatric disorders to the attention of a general neural modeling readership, hoping that this may contribute to further progress in this new and challenging research field.

A major objective of neural modeling is to study the relation between the 'microscopic' features of neural networks of the brain, such as the network's synaptic connectivity and the neuronal firing dynamics, and the 'macroscopic' functional and behavioral phenomena that characterize the network's function. To study brain pathological disorders, one first has to construct a model network which is capable of performing some basic cognitive functions, such as controlling movements or storing and retrieving episodic and semantic memory information. Thereafter, by lesioning the intact network's structural components and disrupting its dynamic mechanisms, the specific neuroanatomical and neurophysiological findings characterizing the pathogenesis of the disease can be modeled, and the resulting changes in the behavior of the network can be examined. Furthermore, it is then possible to search for mechanisms which may counteract the damaging effects of the simulated pathological 'lesions'. Obviously, current neural models greatly simplify both the biological and the cognitive phenomena occurring in brain disorders and are generally constrained in size. Nevertheless, neural models complement traditional methods in substantial and important ways. The pathological changes inflicted on the model network can be controlled precisely

and can be systematically varied over arbitrarily large numbers of experimental 'subjects' and information processing tasks. Further, the experiments simulated are open to detailed inspection in ways that biological systems are not.

This paper reviews the research done in the last decade on neural modeling of psychiatric disorders. What disorders are termed 'psychiatric' and how are they classified? Psychiatric categories are mainly syndromes, where a syndrome is a cluster of symptoms which tend to occur together. In many cases, none of the symptoms manifested in a disease is a defining feature of the illness, and in general psychiatric symptoms are qualitative and difficult to quantify. In most cases, the specific biological causes (termed 'etiological factors') and pathogenic mechanisms of a disorder are not known and the classification is primarily phenomenological. Two main classification systems are used, the DSM which is used primarily in the USA, and the ICD system which has adopted the main principles of the DSM (for an in-depth discussion of classification in psychiatry see [Pichot, 1994]). These diagnostic systems contain the same basic categories, including mental retardation, disorders of childhood, mental disorders, adjustment disorders, personality disorders and 'other' disorders. Mental retardation denotes an impairment of intellectual functioning that is present continuously from early life. The mental disorders category holds the major psychiatric illnesses such as schizophrenia and major depression. These disorders consist of severe behavioral or psychological abnormalities with a recognizable onset after a period of normal functioning. Less severe disorders, which occur in relation to stressful events are termed adjustment disorders, while predispositions to behave in a certain abnormal way are grouped in the category of personality disorders. Although not part of the classification systems, the terms psychosis and neurosis are widely used. Psychoses is a broad term for the more severe form of psychiatric disorders such as schizophrenia, where delusions and hallucinations may occur and the patients may lack insight and do not recognize that they are unwell. Neuroses denote mental disorders in which insight is maintained and delusions and hallucinations are absent.

Most of the work reviewed here is concerned with syndromes which belong to the mental disorders category. The main goal of this paper is to expose the reader to the pertaining

clinical phenomena and the hypothesized neural and synaptic pathological alterations, and to refer the reader to more detailed reviews of these issues in the clinical literature. The current modeling studies are described on a conceptual level; the technical details concerning the formulation of the various models are mostly omitted, as they basically employ standard attractor and feed-forward networks. It is hoped that the interested reader may be 'tempted' to find more about those details in the original papers.

Concentrating on neural models of psychiatric disorders, we shall not discuss related work concerning models of neurological disorders and neuropsychological dysfunction. Models of neurological disorders include, for example, models of epilepsy ([Wong *et al.*, 1986; Mehta *et al.*, 1993], phantom limbs [Spitzer *et al.*,], cortical reorganization after stroke [Sutton *et al.*, 1994; Armentrout *et al.*, 1994] and frontal lobe syndromes [Levine & Prueitt, 1989; Dehaene & Changeux, 1989; Dehaene & Changeux, 1991]. Models of neuropsychological disturbances, many of them developed following the work of the PDP group [McClelland & Rumelhart, 1986a], include models of memory dysfunction [Wood, 1978; Anderson, 1983], disconnection syndromes [Gordon, 1982], amnesia [McClelland & Rumelhart, 1986b], prosopagnosia [Virasoro, 1988], agnosia [Farah & McClelland, 1991], aphasia [Martin *et al.*, ; Plaut & Shallice, 1993b], dyslexia [Reggia & Berndt, 1986; Patterson *et al.*, 1990; Hinton & Shallice, 1991; Plaut & Shallice, 1993a], attentional deficits [Cohen *et al.*, 1994; Humphreys *et al.*, 1992], motivation [Grossberg, 1984] and automaticity [Cohen *et al.*, 1992] (see [Reggia *et al.*, 1994] for a review).

The rest of this paper is organized as follows. In the next section we review the work done on modeling memory deterioration in Alzheimer's disease, the main cause of dementia. In Section 3 we describe the work done on modeling various symptoms and cognitive dysfunctions typically manifested in schizophrenia. In Section 4 we describe models which address some clinical aspects of manic-depressive disorder, major depression and paranoid disorders. Finally, the current status and future prospects of neural modeling of psychiatric disorders is discussed in Section 5.

2 Alzheimer's disease

Alzheimer's disease (AD) is the most common dementing illness. Dementia is a syndrome characterized by global dysfunction. Its essential feature is broad-based intellectual decline, representing a significant decline from previous levels of functioning and causing significant impairment in social and occupational functioning. Although AD is characterized by the development of multiple cognitive deficits manifested by disturbances in language, motor and executive functions, a major clinical hallmark of the disease is memory impairment, which manifests itself as an inability both to recall previously learned knowledge and to learn new information. Thinking is impaired, with poverty of associations, frequent perseverations and 'intellectual inflexibility'. Loss of interest and initiative are also relatively early signs, accompanied by distractibility and fatigue. As the disease proceeds, impaired judgment and impulse control are often manifest, together with marked personality shifts. Paranoid ideation and persecutory delusions (described in detail further on) are frequent psychotic features of dementia, and may be present in almost half of AD patients as the disease progresses.

The clinical course of AD is usually characterized by gradual deterioration, although both slow and rapidly progressive forms have been reported, exhibiting a large variation in the rate of AD progression. The diagnosis of AD is traditionally based on the presence of specific microscopic abnormalities such as neurofibrillary tangles and senile plaques in brain tissue (which in small numbers are found also in normal aging). A confirmed diagnosis of AD is usually made only after autopsies, or, less frequently, after brain biopsy. Most cases of AD are accompanied by ventricular enlargement and generalized cortical atrophy, which is most prominent in frontal and temporal areas. Although considerable progress has been made recently in understanding some neurobiological features of AD, its causes and pathogenesis are still generally unknown (the interested reader is referred to [Katzman, 1986; Price *et al.*, 1993] for extensive reviews of AD).

In the following we review two different neural models of the pathogenesis of AD. The first concentrates on studying the possible role of synaptic deletion and compensation, while

the second investigates the role of synaptic runaway.

2.1 Synaptic Deletion and Compensation

Recent neuroanatomical investigations have repeatedly demonstrated that the progress of AD is accompanied by considerable synaptic changes, including synaptic deletion and compensation. While *synaptic deletion* is manifested in a reduction of the number of synapses per unit of cortical volume, a concomitant increase in the size of the remaining synapses has also been observed and is referred to as *synaptic compensation*.

In light of the major place that memory impairment occupies among the clinical manifestations of AD, and following the abundance of data about neuronal and synaptic degenerative changes that occur as the disease progresses, Alzheimer's disease seems an 'ideal' candidate for neural modeling. More specifically, [Horn *et al.*, 1993] were interested in studying how the interplay between synaptic deletion and compensation determines the observed patterns of memory deterioration, and what strategies of increased synaptic efficacy could best maintain memory capacities in face of synaptic deletion.

Investigating these synaptic changes in a Hopfield-like attractor neural network model of associative memory [Tsodyks & Feigel'man, 1988], [Horn *et al.*, 1993] have shown that the deterioration of memory retrieval due to synaptic deletion can be much delayed by strengthening the remaining synaptic weights by a uniform compensatory factor defined by

$$W_{ij}{}^{new} = c \cdot W_{ij}{}^{old} \ , \qquad c = 1 + \frac{dk}{1-d} \ , \tag{1}$$

where W_{ij} is the connection weight from neuron j to i, c is the compensation factor, d denotes the level of (random) synaptic deletion and k is a 'compensation-strategy' parameter whose value is a function of d. By using different dependencies of k on d it is possible to define various compensation strategies which can account for the observed variation in the severity and progression rate of AD. These results explain the specific patterns of cognitive decline observed in various clinically-defined subgroups of AD patients, and have led to the formulation of a new hypothesis accounting for the appearance of parkinsonian symptoms in AD patients [Horn & Ruppin, 1992]. The work of [Horn *et al.*, 1993] is however limited

in two important ways: First, since a prescribed synaptic memory matrix was used, only memory retrieval could be addressed. Second, the synaptic compensation dependencies employed were realized in a 'global' manner which is biologically unrealistic.

The first limitation was addressed in [Ruppin & Reggia, 1995]. Using a simple, activity-dependent Hebbian synaptic storage scheme to model memory acquisition in the framework of an attractor neural network model, [Ruppin & Reggia, 1995] examined a recent claim that neural models cannot account for more detailed aspects of memory impairment, such as the relative sparing of remote versus recent memories observed in Alzheimer's patients [Carrie, 1993]. The model exhibits differential sparing of remote versus recent memories, accounts for the experimentally observed temporal gradient of memory decline, and shows that neural models can account for a large variety of experimental phenomena characterizing memory degradation in Alzheimer patients. Specific testable predictions have been generated, concerning the relation between the neuroanatomical degenerative findings and the clinical manifestations of Alzheimer disease.

The biological appeal of the uniform synaptic compensatory regimes studied by [Horn *et al.*, 1993] hinges upon the ability to show that they can actually be realized in a 'local' manner, where each neuron readjusts its synaptic weights only as a function of local information such as its post-synaptic potential. A recent study of such local compensatory mechanisms conducted by [Horn *et al.*, 1995] demonstrates that this is a non-trivial but feasible task. This study has revealed a new dependency between the extent of synaptic changes and the retrieval properties of the network. In contradistinction to the case of global compensatory strategies, the network's performance does not only depend on the current magnitudes of deletion and compensation, but it also depends on the precise rates at which these processes progress. That is, the performance of the network is history-dependent. This dependency provides a new explanation to the rather puzzling broad variability in structural indicators of damage observed in AD patients having approximately similar levels of cognitive function.

2.2 Synaptic Runaway

While the studies reviewed above have sought to explain memory degradation in AD as a failure of synaptic compensatory responses to account for ongoing accelerated synaptic deletion, a different approach has been taken by [Hasselmo *et al.*, 1992; Hasselmo, 1993; Hasselmo, 1994], studying *runaway synaptic modification* in both attractor and feed-forward networks. Runaway synaptic modification denotes a pathological exponential growth of synaptic connections, that may occur due to interference of previously stored patterns in the storage of new patterns. This interference occurs because, when a new memory pattern is being stored in the network, the resulting network activity is not only guided by the new pattern but also by all the previous memory patterns which are engraved in the synaptic matrix. Thus, previously memorized patterns tend to bias the activation during new storage in 'their direction'. This inherent reenforcement may lead to exponential synaptic growth and to a pathological increase in the number of synapses.

One possible way to prevent synaptic runaway is to assume that the strength of the external projections via which new patterns are stored in the network is sufficiently strong to overcome the interference of other memories (this assumption does not necessarily imply the use of strong external fields that 'clamp' the activation in the network - see [Ruppin & Reggia, 1995]). Another alternative, raised by Hasselmo and his coworkers, is that runaway synaptic modification can be inhibited by suppression of internal synaptic connections (synapses between neurons belonging to the same cortical module) during learning.

What is the hypothesized role of synaptic runaway in the pathogenesis of AD? Hasselmo's analysis shows that there is a critical storage capacity beyond which interference during learning cannot be prevented and synaptic runaway is unavoidable. Several factors can lead to the initiation of synaptic runaway, such as a decrease in the level of cortical inhibition, reduced synaptic decay, and excess memory storage. Once synaptic runaway occurs, it is claimed that its increased metabolic demands or excitotoxic effects could be sufficiently severe to cause neuronal degeneration, parallel to that found in AD. Furthermore, Hasselmo's work provides a theoretical framework for describing the specific distribution of neuronal degeneration observed in AD, where entorhinal regions lacking suppression of

internal synaptic transmission are more markedly damaged than other cortical regions.

Hasselmo's theory has been inspired by experimental work [Hasselmo & Bower, 1992; Hasselmo & Barkai, 1992] that provides evidence that acetylcholine selectively suppresses excitatory synaptic transmission at the internal synapses, while allowing external afferent synaptic transmission (i.e., projections from neurons belonging to other modules) to operate at full strength. Accordingly, it is claimed that the loss of cholinergic innervation in AD may underlie the initiation of runaway synaptic modification, and that sprouting of cholinergic innervation observed in the dentate gyrus during AD reflects attempts to arrest the progress of synaptic runaway.

Hasselmo's work is an excellent example of research which combines experimental physiological studies with computational modeling. It demonstrates how a computational model can raise a quandary (how are patterns actually stored without being accompanied by synaptic runaway?) which motivates an experimental study (the differential effects of acetylcholine on internal and external synapses). Moreover, the theoretical solution to this question gives rise to further hypotheses concerning the possible consequences of a disruption of the newly revealed computational mechanism (i.e., the role of synaptic runaway in the pathogenesis of AD).

The review of neural models of Alzheimer's disease will not be complete without mentioning several new models of memory function that were published recently. These models, whose detailed description is beyond the scope of this paper, present interesting attempts to provide a general framework of hippocampal-cortical interaction which can account for memory consolidation and retrieval [Alvarez & Squire, 1994; McClelland *et al.*, 1994; O'Reilly & McClelland, 1994; Tsukuda, 1994]. Such integrative models of normal memory function will probably trigger further investigations of the functional significance of their disruption.

3 Schizophrenia

To the clinician and researcher alike, schizophrenia remains perhaps the most enigmatic of the psychiatric disorders. It is a clinically heterogeneous disorder with a broad spectrum of manifestations, and yet it defies sub-classification. Until today, our diagnostic criteria still rely on observations of psychiatric phenomenology. The disease affects a broad range of cognitive and emotional systems. The symptoms are diverse, and include hallucinations, delusions, disorganized speech and behavior, loss of fluency of thought and speech, impaired attention, abnormalities in the expression and observation of emotion, and loss of volition and drive. None of these symptoms is a defining feature of the illness. However, psychosis (that is, a breakdown with reality manifested in hallucinations, delusions, or impaired thought processes) and a significant deterioration in functioning have become central to the diagnosis of schizophrenia. The course of the illness tends to be marked by exacerbations and remissions, but there is rarely a full return to pre-morbid functioning levels. The persistence of the impairment gives the disease a 'dementia-like' quality, but it differs from classic dementia in that most schizophrenic patients stabilize at moderate levels of cognitive impairment, and the disease does not have a progressive downhill course leading eventually to death. (The interested reader is referred to [Roberts, 1990; Waddington, 1993; Carpenter & Buchanan, 1994; Andreasen, 1994] for excellent recent reviews of the pathological findings and clinical manifestations observed in schizophrenia).

In an attempt to conceptually organize and simplify the clinical presentation of schizophrenia, its clinical symptoms are typically divided into two categories: 'positive' and 'negative'. The positive symptoms are 'productive', that is, exaggerations or distortions of functions that are normally present, and include delusions, hallucinations, thought and speech disorders and bizarre behavior. The negative symptoms represent loss or diminution of functions, and include poverty of thought and speech, emotional bluntness, and impaired attention. Another approach classifies schizophrenic symptoms into three categories: psychotic symptoms (delusions and hallucinations), dissociative thought disorders and impaired

attention, and negative symptoms. Both systems of categorization have gained support from factor analyses studies.

The pathogenesis of schizophrenia is yet unknown. A few theories have been raised, based on neuropathological observations, the actions of anti-psychotic medications, and ideas about the relation between brain and behavior. Perhaps the most enduring biochemical explanation of the pathophysiology of schizophrenia is the dopamine hypothesis, which currently postulates the coexistence of hypodopaminergic activity in the mesocortical system, resulting in negative symptoms, and hyper-dopaminergic activity in the mesolimbic system, resulting in positive symptoms. Structural and functional imaging and neuroanatomical postmortem studies are providing converging evidence of the involvement of specific brain regions in schizophrenia, such as the prefrontal areas, temporal lobes and the temporo-limbic circuitry, and subcortical and midline circuitry. Integrative pathophysiological hypotheses have attempted to explain schizophrenic symptoms in terms of biochemical and neuroanatomical alterations in specific brain circuits, but at present no single explanatory mechanism has prevailed (a few of the most prominent of these theories were presented in [Stevens, 1973; Weinberger, 1987; Carlsson & Carlsson, 1990; Stevens, 1992]).

Neural modeling efforts of schizophrenia have also taken two main paths, perhaps reflecting the view of schizophrenia as composed of positive symptoms that arise due to temporo-frontal pathology, and negative symptoms that are a result of prefrontal abnormalities. This is true both with regard to the symptoms modeled, and, to the models employed. The first avenue, pioneered by Hoffman, has concentrated on modeling schizophrenic positive symptoms in the framework of an associative memory attractor network [Hoffman, 1987; Hoffman & Dobscha, 1989]. This work has 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 it is locked in dysfunctional patterns of activity. In another framework, of feed-forward layered networks employing back-propagation learning, Cohen and Servan-Schreiber have provided a detailed computational account of how schizophrenic functional deficits can arise from neuromodulatory effects of dopamine [Servan-Schreiber *et al.*, 1990; Cohen & Servan-Schreiber, 1992a].

3.1 Modeling Positive Symptoms with Attractor Networks

3.1.1 Synaptic Deletion, Memory Overload and Parasitic Attractors

A few formal models of information processing breakdown have been presented earlier (e.g., [Callaway, 1970; Broadbent, 1971; Joseph *et al.*, 1979; Callaway & Naghdi, 1982]), but the publication of Hoffman's [1987] paper probably marks the beginning of 'the era of neural modeling' of schizophrenia. In this paper, Hoffman describes how pathological alterations in a Hopfield attractor neural network can lead to the formation of *parasitic attractors*, whose cognitive and perceptual manifestations may play an important role in the emergence of schizophrenic delusions and hallucinations. These parasitic states are spurious states that are generated when the network becomes 'overloaded', i.e., its memory capacity is exceeded and catastrophic breakdown occurs [Amit, 1989]. Such memory overload presumably occurs in the brain of schizophrenics as a result of neurodegenerative changes, or as a result of selective attention deficits.

Delusions are common abnormalities of thought among schizophrenics. They are defined as idiosyncratic, false beliefs, that are unshakable. Such beliefs cannot be disproved by logical arguments to the contrary, and they are not shared by members of the same culture. Typical delusional themes of schizophrenic patients consist of externally imposed influences (thought insertion and thought broadcasting), grandiose delusions (a belief that one has unusual talents or an identity of a famous person), erotomania (in which the patient believes that a famous person is in love with him), and persecutory delusions (being a target of malevolent action). The inescapability of delusions, and their being spontaneously invoked at times by seemingly irrelevant experiences, have led Hoffman to the idea that they can be conceived as parasitic attractor states which have broad and 'deep' basins of attraction. Hoffman also proposed that hallucinations have a similar linkage to parasitic states. A hallucination is a perceptual disturbance which consists of a perception in the absence of an external stimulus. The most common hallucinations in schizophrenics are auditory hallucinations, where typically the voices perceived appear to be coming from outside and making a running commentary on the patient's behavior. Other types of hallucinations such

as visual and olfactory hallucinations are less common but may also occur. The primary feature differentiating schizophrenic hallucinations from normal imagery is that the former are experienced as alien and out of control, properties which are reminiscent of a parasitic state which is not a learnt, familiar, memory state.

The linkage between parasitic states and delusions and hallucinations stems from the alien nature of the latter and their tendency to be repetitive. Building upon the basic linkage between parasitic states and positive symptoms, [Hoffman & Dobscha, 1989] have presented a detailed simulation study that examined the hypothesis that the onset of schizophrenia (usually marked by a psychotic crisis and positive symptoms) is triggered by progressive elimination of synapses in the prefrontal cortex. In accordance with this hypothesis, the pathological excess of synaptic pruning reflects a normal developmental synaptic elimination process that fails to arrest in time and proceeds too far (interestingly, the typical onset period of schizophrenia is during late adolescence, when synaptic pruning supposedly reaches it peak). Studying this hypothesis in a 2-D associative memory attractor neural network, prefrontal synaptic pruning is modeled as a process of random synaptic deletion that tends to damage weak and distal synaptic connections more than strong and proximal ones, such that synapse W_{ij} from neuron j to neuron i is pruned only if

$$|W_{ij}| < p \cdot D_{ij} \ , \tag{2}$$

where D_{ij} denotes the distance on the lattice between neurons i and j and p is the pruning factor. This type of spatial-selective damage was found to lead to two kinds of behavior in the network that may have interesting parallels in schizophrenic symptomatology: 1. 'Functional fragmentation' - denoting patches of convergence to different memories in distinct regions of the network. 2. Spatially organized 'parasitic foci' - denoting patches of the network that tend to lock into some non-memory activation patterns regardless of initial input cues applied to the network. [Hoffman & Dobscha, 1989] suggest that the observed functional fragmentation models the 'contamination response' that is specific to schizophrenics, i.e., the fusion of multiple distinct gestalts presented in each image of the Rorschach personality test.

In addition to the intuitive notion that schizophrenic delusions and hallucinations typically arise in a spontaneous and repetitive manner, what other characteristics of ill-formed attractor states can be thought of as linked to the pathogenesis and manifestations of schizophrenic positive symptoms? [Hoffman & McGlashan, 1993] have recently provided a detailed account of the possible role of parasitic foci in the formation of schizophrenic positive symptoms, suggesting 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 attractor. 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. 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 & Arpia, 1994] have recently proposed that due to pathological changes the brain tends to settle in certain attractors which obtain a psychotic 'attunement'.

3.1.2 Synaptic Compensation and Spontaneous, Biased Memory Activation

Following the work of Hoffman and his colleagues, [Horn & Ruppin, 1995] have examined a recent theory of Stevens [1992] in the framework of an attractor neural network model. As summarized in [Stevens, 1992], the wealth of data gathered concerning the pathophysiology of schizophrenia suggests that there are atrophic changes in temporal lobe regions 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 temporal lobe neurons, including the frontal cortex. 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 degenerating temporal neurons.

To study the functional implications of Stevens' hypothesis, [Horn & Ruppin, 1995] modeled a frontal module as an associative memory neural network receiving its inputs from degenerating temporal projections and undergoing reactive synaptic regeneration. In this model, it is shown that while preserving memory performance, compensatory synaptic regenerative changes modeling those proposed by Stevens may lead to adverse, spontaneous activation of stored patterns. When *spontaneous retrieval* emerges, the incorporation of Hebbian activity-dependent synaptic changes leads to a *biased* retrieval distribution that is strongly dominated by a single memory pattern [Ruppin *et al.*, 1995]. The hypothesized activity-dependent pathological synaptic changes are modeled via the Hebbian rule

$$W_{ij}{}^{new} = W_{ij}{}^{old} + \gamma(\bar{S}_i - a)(\bar{S}_j - a) \ , \tag{3}$$

where $\bar{S}$ is 1 (0) only if the neuron has been consecutively firing (quiescent) for some recent time period, a is the coding level, and γ is a constant. The formation of biased, spontaneous retrieval is shown to require the concomitant occurrence of both degenerative changes in the external input (temporal) fibers and regenerative activity-dependent Hebbian changes in the intra-modular (frontal) synaptic connections.

A few important characteristics of positive symptoms are reflected in the behavior of the network: 1. The emergence of spontaneous, non-homogeneous retrieval is a self-limiting phenomenon; eventually, a global, spurious, attractor is formed. The formation of such a cognitively meaningless spurious attractor, accompanied by a decrease in the size of basins of attraction of the memory patterns, may lead to the emergence of deficit, negative symptoms. This parallels the clinical observation that as schizophrenia progresses positive symptoms tend to wane, while negative symptoms are enhanced. 2. When the network converges to a memory pattern that dominates the output in the spontaneous-retrieval scenario, it has increased tendency to remain in this state for a much longer time than in its normal functioning state, in accordance with the persistence of positive symptoms. 3. The model

points to the possibility that maintenance therapy may have an important role not only in preventing the recurrence of positive symptoms, but also in slowing down the progression of the disease, by arresting the pathological evolution of the synaptic memory matrix. 4. In its spontaneous retrieval mode, the network may also converge to mixed retrieval states, which have some similarity to a few patterns concomittantly. 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. 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.

The occurrence of autonomous, biased, memory activation parallels Hoffman's concept of parasitic foci. However, some significant points of difference should be noted. First, while Hoffman's work concentrates on modeling the effects of synaptic degenerative changes, [Horn & Ruppin, 1995; Ruppin *et al.*, 1995] 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 [Ruppin *et al.*, 1995] 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 is preserved in [Ruppin *et al.*, 1995] until late stages in the evolution of biased retrieval. The latter difference is important since memory is generally preserved in the early stages of schizophrenia. Finally, recent cognitive studies show that 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 large basins of attraction (such as the biased attractors described in [Ruppin *et al.*, 1995] have), and are not simply fragments of independent activity as in [Hoffman & Dobscha, 1989]. Interestingly, [Ruppin *et al.*, 1995]'s account provides a neural 'correlate' of the widely held notion that delusions and hallucinations are adaptive responses to preexisting disorganization as part of a compensatory 'defense' mechanism.

The generation of spontaneous pattern activation following neural damage that alters the input/internal synaptic balance is a quite general phenomena, as recently demonstrated in [Thaler, 1995].

In addition to the obvious simplifications involved in constructing a neural model of higher brain functions, the work presented in [Horn & Ruppin, 1995; Ruppin *et al.*, 1995] is lacking in two specific aspects: 1. While Stevens' theory involves changes occurring in numerous cortical and subcortical structures, we focus on only a simple, canonical, computational model of Stevens' hypothesis that includes only a small subset of the brain structures involved. 2. Only a single frontal network, or module, is examined: 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. Few possible solutions to this quandary may be studied in the future, as described in detail in [Ruppin *et al.*, 1995].

3.2 Modeling Cognitive Functions With Layered Networks

A different approach, both with regard to the phenomena studied and the models employed, has been taken by Cohen and Servan-Schreiber. Building upon their work on modeling the neuromodulatory effects of catecholamines on information processing [Servan-Schreiber *et al.*, 1990; Servan-Schreiber & Cohen, 1992], they have presented a comprehensive modeling study of the performance of normal subjects and schizophrenics in three attentional and language processing tasks [Cohen & Servan-Schreiber, 1992a; Cohen & Servan-Schreiber, 1992c; Cohen & Servan-Schreiber, 1993]. These tasks are important indices of cognitive dysfunction in schizophrenia, and are related to schizophrenic negative symptoms. Their modeling has enabled a detailed quantitative investigation, which is not confined to the qualitative realm of positive symptoms.

The tasks modeled were the Stroop test, which examines the ability to respond to one set of stimuli even when other, more compelling stimuli are available, the continuous performance test, where the subjects' task is to detect target stimuli within a stream of stimuli

including distractors, and a lexical ambiguity task which examines the temporal range over which the subjects can use context to choose the appropriate meaning of ambiguous words. The Stroop task was modeled using a layered feed forward network while the other tests required the use of a three-layered network with recurrent connections, to encompass a temporal dimension. In all tasks, a back propagation algorithm was used to train the networks to simulate normal performance. Although each task was modeled by a network designed specifically for that task, the networks used rely on similar information processing principles and share a common module for representing context, which is identified by the authors with the prefrontal cortex. Neuronal firing is governed by a sigmoidal function $S(h) = 1/(1 + \exp\{-g \cdot h + b\})$, where h is the input field of the neuron, g the gain of the sigmoid and b its bias. The hypothesized neuromodulatory effects of dopamine on information processing (which may play a major role in the pathogenesis of schizophrenia, as described above) were modeled as a global change of the gain g. The simulations performed demonstrate that a change in the gain of neurons in the context module can quantitatively account for the differences between normal and schizophrenic performance in the tasks examined. As quantitative estimates of dopamine deficits in schizophrenia are currently absent, the gain value was first chosen to ensure adequate performance in normals, and the hypothesized neuromodulatory changes in schizophrenia were modeled using the same gain value in all networks.

Cohen and Servan-Schreiber review a large amount of data that places their studies in an interesting perspective: Due to abnormalities in the activity of the mesocortical dopaminergic system, the function of the frontal cortex in schizophrenia is disturbed, resulting in decreased performance of schizophrenics in behavioral tasks relying on the use of context. As in any other current model of schizophrenia, the work focuses on a single disturbance (in this case, concerning context representation) while there are probably several disturbances underlining the disruption of cognition in schizophrenics. The study addresses a small set of tasks, but these may represent important cognitive correlates of schizophrenic negative symptoms. Obviously, the affective (emotional) aspects of negative symptoms cannot be addressed within the neural modeling paradigms currently available. Cohen and Servan-

Schreiber's work on modeling dopamine effects and schizophrenic deficits obviously leaves many questions open, which has recently led to a vigorous discussion of their theory and its implications (primarily, the relation between the neuromodulatory 'gain' and dopamine) in the psychiatric literature [Jobe *et al.*, 1994; Cohen & Servan-Schreiber, 1994]. Based on the work of [Servan-Schreiber, 1990] on simulating human performance in a choice-reaction time task (Eriksen task), [Callaway *et al.*, 1994] have recently reanalyzed data from a similar task performed by subjects which have been under the influence of various drugs. The authors maintain that patterns of performance observed in the data are in accordance with those predicted by Servan-Schreiber's model, when considering possible drug neuromodulatory effects on the gain and bias of units in different layers. They claim that "neural network models offer a better chance of rescuing the study of human psychologic responses to drugs than anything else currently available" [Callaway *et al.*, 1994].

Although the model presented in [Cohen & Servan-Schreiber, 1992a] has not addressed schizophrenic positive symptoms, the authors have pointed out that these may be studied in a similar modeling framework. This issue is precisely the goal of a recent study by [Hoffman *et al.*, 1994]. Aiming to provide a quantitative description of the pathogenesis of auditory hallucinations, [Hoffman *et al.*, 1994] have studied the hypothesis that hallucinated 'voices' arise from altered verbal working memory. The model consists of a recurrent layered neural network with a temporary storage layer (with similar architecture to that used in [Cohen & Servan-Schreiber, 1992a]) and uses backpropagation to learn a speech perception task. In this task, sequences of randomly coded input words (referred to as 'phonetic') are translated into sequences of outputs (referred to as 'semantic') in a semantic feature space (actual acoustic data was not used). In parallel to the simulation studies, an experimental study of speech perception in schizophrenics and normal controls was conducted. In accordance with the now 'standard' paradigm in these type of combined experimental/modeling studies, the network's architectural and dynamical parameters were first tuned to model the performance of normal subjects. Thereafter, various alterations of network connectivity and dynamics were systematically studied and the resulting network performance patterns were compared with the experimental findings. In addition to several indices of speech per-

ception impairment that were traced, 'auditory hallucinations' were defined as the detection of words by the network (i.e., a significant activation of the output layer) in the absence of any phonetic input.

Reviewing the relevant neuroanatomical and neurophysiological data, three rivalry models of the pathogenesis of schizophrenia were examined; neuromodulatory alterations (modifying the bias and gain of the sigmoid function governing the function of hidden layer neurons), neuroanatomical alterations (various degrees of pruning the connections from the temporary ('working memory') layer to the hidden layer), and a combination of both. The authors used an interesting method to assess the three rivalry pathogenesis models studied: Dividing the various indices of speech perception they studied into two groups, they optimized the parameters of each model to obtain near perfect matching with the experimental results pertaining to one group of indices. They then compared how each model matched the experimental results pertaining to the second group of indices. Using this method, the model with combined neuroanatomical/neuromodulatory disturbances demonstrated clear superiority as a predictor of speech disturbances in schizophrenics. Several interesting insights and predictions have been generated: 1. Schizophrenics with auditory hallucinations should have significantly more severe speech perception abnormalities relative to non-hallucinating patients. 2. Several types of alterations can lead to auditory hallucinations, but the combined anatomical/modulatory model is the most likely one. 3. Drugs altering the response profile of neurons may be effective even when the primary pathology is neuroanatomical rather than neuromodulatory. 4. With severe neuroanatomical damage, perceptual function must be sacrificed in order to reduce hallucinations. Interestingly, this finding closely corresponds to that of [Ruppin *et al.*, 1995] in an entirely different framework.

4 Other disorders

While a considerable number of the neural modeling studies done in recent years have focused on cognitive and perceptual disturbances in schizophrenia, many other symptoms and disorders have also been studied, as described in this section. There is no single 'unifying theme' which underlies the grouping together of these studies; the disturbances investigated are modeled at various levels of description using different basic neural models, and the organization of this section is quite arbitrary.

4.1 Paranoid and Dissociative disorders

Perhaps the topic most closely related to schizophrenia modeling is the study of paranoid processes. Paranoia is a tendency to develop suspicions and ambitions that gradually progress to 'systematized' delusions of persecution and grandeur. This chronic and unremitting system of delusions encompasses a broad set of false ideas that are connected by a common theme and are rigidly adhered to despite all contradictory evidence. The quality of the delusions is non-bizarre, i.e., possible even though implausible, and quite characteristically, facts are reinterpreted to fit the delusions rather than vice versa. A paranoid disorder is distinguished from schizophrenia by the absence of hallucinations and other psychotic features, by the relative preservation of personality, and by lack of functional and occupational impairment outside the areas of life directly involved in the delusional system.

[Vinogradov *et al.*, 1992] have recently described a model of paranoid processes within the framework of spreading activation networks. Motivated by high-level psychological models of semantic memory and associations, each computational unit represents a distinct cognitive item, and the links between units represent associations. The authors propose that paranoia gradually forms in a process where initial suspicions consolidate into a delusional system. In this process, associations are constructed among temporally contiguous perceptions in an excessive manner, and are assigned an idiosyncratic meaning of malevolent motives or persecution by others. The authors suggest that this process can be modeled by a spreading activation network whose connectivity and dynamical parameters are altered, reflecting a 'hyper-associative' state. Relying on the work of [Shrager *et al.*, 1987],

[Vinogradov *et al.*, 1992] describe various phase transitions that the network undergoes when its structural and dynamical parameters are changed. There are three possible phases characterized by different sizes of activity clusters, where large and persistent clusters represent a delusional system. The work of [Vinogradov *et al.*, 1992] concentrates more on the conceptual level and not on a neural network realization. As such, it does not suggest a link between specific neuropathological changes and paranoid symptoms, but presents a formal framework that may be relevant for studying the pertaining psychological data.

Just as the modeling of paranoid disorders seems to require some appropriate way of representing a 'paranoid system', the modeling of another psychiatric subclass, that of dissociative disorders, requires some preliminary understanding of the concept of consciousness. Dissociation is a disturbance in the integration of identity, memory, or consciousness. Information is unavailable to consciousness and yet indicates its presence, as for example in hypnotized subjects. Dissociative disorders include rare but popularly celebrated disorders such as multiple-personality disorder (often confused in the public mind with schizophrenia), fuga, and psychogenic, post-traumatic amnesia. Taking a first step towards the development of a neural model of dissociative disorders, [Li & Spiegel, 1992] have recently presented a thought-provoking discussion of the possibility of modeling these disorders in the framework of error-propagation feed-forward networks. However, it is not clear how to proceed and study such a specific model of dissociative states.

4.2 Attractor Neural Network Models of Manic-Depressive Disorder and Delirium

In addition to modeling memory deterioration and schizophrenic positive symptoms, attractor neural networks have been considered as a framework for modeling a few cognitive manifestations of manic-depressive disorder . Manic-depressive disorder is an affective disorder which includes patients with mania and depression or mania only. The manic bouts are characterized by a distinctly elevated, expansive or irritable mood, accompanied by 'hyperactivity' symptoms such as decreased need for sleep, pressure to keep talking, 'racing' of thoughts and inflated self-esteem. These manifestations are accompanied by

marked social impairment, and frequently, also by psychotic symptoms such as delusions and hallucinations.

As described in Section 3, [Hoffman, 1987; Hoffman, 1992] has proposed that schizophrenic parasitic attractors are formed due to pathological processes which result in 'memory overload'. As a result, previously separate basins of attraction may unify, leading to spontaneous transitions from one memory pattern to another. Hoffman saw this phenomenon as a possible correlate of 'loosening of associations', a characteristic disturbance of schizophrenic discourse resulting in disorganized speech and in failure to maintain a coherent theme. In the framework of this conceptual metaphor, Hoffman proposed that manic 'hyperactivity' may arise not as a result of structural damage to the network, but due to an increase in the noise levels resulting in enhanced rate of transition between attractors. Hoffman's views were inspired by the lack of evidence for widespread neuroanatomic damage in manic-depressive disorder, by the success of lithium drugs which alter neural metabolic pathways and possibly their firing dynamics, and by discourse studies that show that the basic structure of manic discourse is intact. Obviously, the neural representation of a 'thread of thoughts' as a series of transitions between attractor states might be a gross simplification, but perhaps the principles of structural versus functional damage embodied in the metaphors described above would remain of relevance also in more developed descriptions.

In a similar spirit to Hoffman's ideas about the possible role of noise increase in generating 'hyperactivity' symptoms in mania, Wang and Ross have studied the functional effects of several variations in the structure and dynamics of attractor neural networks, and proposed that the latter may have a few cognitive correlates. [Wang & Ross, 1990] investigated how external projections effect the retrieval acuity of an attractor neural network, and suggested that such mechanisms may play a role in distraction and concentration. [Wang & Ross, 1991] have shown that a dynamically tuned neuronal threshold can significantly improve retrieval performance, and suggest that such dynamical threshold tuning may play part in mechanisms of selective attention.

The possible role of threshold variation in the pathogenesis of delirium has been recently

explored in a neural modeling study by [Avni *et al.*, 1995]. Delirium is characterized by a transient impairment of a wide range of cognitive functions due to a diffuse derangement in cerebral metabolism. It has an abrupt onset and a relatively brief duration marked by a fluctuating course. Impairment of consciousness and reduced awareness to the environment are hallmarks of the disorder, together with cognitive dysfunction which may include poor memory, slowness of thinking, inconsistent responses and difficulty in concentrating. Almost any process that causes a diffuse disruption of brain homeostasis, such as fluid and electrolyte disturbances, drugs and infections, may cause delirium. Little is known on the pathogenesis of delirium and the current main paradigm is that it is a common pathway for a variety of metabolic pathological processes. [Avni *et al.*, 1995] have examined the hypothesis that variations in the neural threshold underlie some memory-related cognitive disturbances in delirium. The attractor neural network model they used incorporates two sets of connections; Hebbian connections storing memory patterns, and randomly-weighted connections. Depending on the values of the neuronal threshold and synaptic connectivity parameters, the network may either converge to a stable state or wander through its state space in a seemingly chaotic manner. The transition from the region of single stable states with near perfect retrieval to unstable end states is sharp, and is accompanied by a 'syndrom' of poor memory retrieval, slower retrieval time, instability and inconsistency of end states, and storage disturbances, all which are typical characteristics of memory and cognitive functioning in delirium. Interestingly, similar unstable end states were found in the model at high and low levels of neuronal activity, offering some insight as to how can both excitatory and inhibitory etiological factors (such as low or high levels of some electrolytes) cause delirium. This model also offers an explanation as to why infants and elderly patients are more prone to delirium than subjects from other age groups.

4.3 Major Depression

Major depression, the most prevalent affective disorder, is characterized by low mood and symptoms like loss of interest, psychomotor retardation, fatigue, sleeplessness, impaired concentration and suicidal ideation. The difficulty of representing the symptoms above in

neural network models has restricted current modeling attempts to some cognitive aspects of major depression. Past work related to major depression has concentrated primarily on modeling learned helplessness, an experimental psychological model of depression, in an adaptive resonance network [Leven, 1992]. A similar modeling framework has been used by [Hestenes, 1992] to model the selection and execution of behavioral plans in manic-depressive patients.

Two new approaches to studying major depression are currently under investigation. The first employs backpropagation feed-forward neural networks for performing functional sub-typing in major depression [Luciano *et al.*, 1994b] (functional subtypes are defined as a subset of symptoms that cluster together when correlated with response to treatment). If successful, such studies could lead to more efficient pharmacological treatment. The goal of a second study is to develop a neural model examining the possible role of the limbic system in depressive disorders [Luciano *et al.*, 1994a], creating 'linking hypotheses' between model variables for brain regional activities and clinical symptoms data. This study has very challenging goals, but it is not clear to what extent can one lump together different patterns of activity in a given neural structure into a single variable. However, the success of such projects may further encourage the development of extensive laboratory and clinical data banks for psychiatric disorders, a necessary step towards developing more quantitative models.

5 Discussion

The main goal of this paper is to introduce the readers to the themes that have been dealt with recently in psychiatric neural modeling. These themes include various approaches to studying the role of synaptic changes in the pathogenesis and clinical manifestations of Alzheimer's disease, the study of spurious attractors as possible neural correlates of schizophrenic positive symptoms, and the ability of feed-forward and recurrent network models to quantitatively model human performance in various cognitive tasks, both in normal subjects and in psychiatric patients. Obviously, the studies reviewed in this paper, summarized in Table 1, represent just a beginning. The models presented here all employ

gross simplifications, both on the level of their structure and dynamics, and on the level of the behavioral phenomenology they seek to mimic. However, it is encouraging to realize that the disruption of simple computational mechanisms can lead to behaviors which correspond to some interesting clinical features of psychiatric disorders.

Since there is still a great deal we do not understand about the workings of the brain, most of the work reviewed in this paper could be questioned and criticized. No one currently knows what are the 'right' models of information processing in the brain, and most probably information is processed in different ways in different regions. The neural representations of the behavioral and clinical phenomenology we seek to study are yet to be discovered, and the literature is abundant with a broad spectrum of pathological findings on different levels of description, which give rise to numerous theories of the pathogenesis of psychiatric disturbances. Even the clinical classification and description of psychiatric disorders is still 'a process in evolution', and most psychiatric entities are still defined on an abstract, qualitative level.

In light of this, one possible stance is that the topics and means of psychiatric neural modeling are currently too loosely constrained, and that we should better defer such research efforts to future times. My view, however, is different. While we should certainly be aware of the current limitations of our work, I think that neural models offer a promising and challenging way to explore various hypotheses concerning the pathogenesis of brain disorders in a computational manner. As put by [Frith, 1991] in a recent commentary on a theory of schizophrenic neuropsychology (by [Gray *et al.*, 1991]), one "would find the circuit diagrams more convincing if the verbal descriptions of how they operate were backed by a computational model".

Hopefully, the work reviewed in this paper demonstrates that neural models are a potentially useful methodological tool for examining the feasibility of theoretical hypotheses within a computational context. They can offer new insights into the experimental data, and may unify previously unrelated observations [McClelland, 1988]. Even the much simplified models reviewed here are sufficiently complicated to generate interesting and non-trivial predictions; the feed-back structure of the models frequently nullifies the validity of the

Disorder	Symptoms	Network	Main themes & Methodology	Studies
Alzheimer	Memory impairment	ANN	Retrieval impairment, Synaptic deletion & compensation. - Analysis & simulations	Horn et al 93, Ruppin & Reggia 95, Horn et al 95
Alzheimer	Memory impairment	ANN & FFN	Storage impairment, Synaptic runaway, Neuromodulation. - Analysis & simulations	Hasselmo 93,94
Schizophrenia	Positive symptoms	ANN	Memory overload, Synaptic Pruning, Parasitic attractors & fragmentation - Simulations	Hoffman 87,92 Hoffman & Dobscha 89
Schizophrenia	Speech disturbances, Auditory hallucinations	RNN	Structural versus dynamical alterations. - Simulations	Hoffman et al 94
Schizophrenia	Cognitive dysfunction	FFN & RNN	Neuromodulation, Prefrontal cortex & context alterations. - Simulations	Cohen & Servan-Schreiber 92,93
Schizophrenia	Positive symptoms	ANN	Spontaneous, biased retrieval. - Analysis & simulations	Horn & Ruppin 95, Ruppin et al 95
Paranoid disorder	Paranoid delusions	SAN	Hyper-associative states. - Simulations	Vinogradov et al 92
Manic-Depressive disorder	Flight of ideas	ANN	Neuromodulation. - Simulations	Hoffman 87
Major Depression	Learned helplessness.	ARN	Frontal filtering. - Simulations	Leven 92
Major Depression	Depressive symptoms.	FFN	Time-series clustering, Physiological modeling. - Simulations	Luciano et al 94
Delirium	Cognitive alterations	ANN	Threshold modulation, Age-dependent synaptic profiles. - Simulations	Avni et al 95
Drug effects	Eriksen test	Layered network	Quantitative fit. - Simulations	Servan-Schreiber 90 Callaway et al 94

Table 1: Neural models of psychiatric disorders: A summary. Abbreviations used above: ANN - attractor neural network FNN - Feed-forward network RNN - Recurrent neural network ARN - Adaptive resonance network SAN - Spreading activation network

'transparency assumption' which states that, apart from the lesioned sub-system, a cognitive system of a brain-damaged patient remains essentially similar to that of a normal patient [Farah & McClelland, 1992]. This makes the study of lesioned models a considerable challenge. Studying neural models of brain disorders also presents an additional way to examine the plausibility of the currently available models of *normal* cognitive functioning, and this 'reverse engineering' task would probably always be an important goal of brain disorder studies. For an interesting discussion of the adequacy of neural models for modeling neuropsychological functions see [Pich & Guidice, 1992].

What are the future challenges and prospects of psychiatric neural modeling? This is difficult to predict, but it may be of interest to mention a few issues in this respect.

1. *Developing neural models of more complex cognitive function:* Current work has concentrated on making use of available neural modeling tools. This has restricted the cognitive phenomena studied to memory-related processes, and to learning relatively simple tasks in a supervised manner. The development and incorporation of more sophisticated neural models is probably an essential step towards capturing more complex phenomena. Promising venues include models of reenforcement learning, multi-modular associative memories, and multi-layered recurrent networks.

2. *Modeling new experimental data:* Recent advancements in several experimental techniques have yielded a number promising developments. Of special interest to neural modelers are the recent developments in techniques that provide information on neural and synaptic degenerative processes. Those include neuroanatomical morphometric and immunochemical methods employed in studying Alzheimer's disease (e.g., [Hansen *et al.*, 1988; DeKosky & Scheff, 1990]), and magnetic resonance spectroscopy which can be used to study neural membrane alterations in Alzheimer and schizophrenia *in vivo* [Pettegrew *et al.*, 1991; Petegrew *et al.*, 1993]. (For a discussion of future implications of these methods to psychiatric modeling see [Hoffman & McGlashan, 1993]). In parallel to these methods which provide knowledge on the 'microscopic' (or, say 'mesoscopic') level, interesting data has been

recently gathered also on the phenomenological, macroscopic level. Novel studies of indirect priming now suggest that semantic associative memory operates at a comparatively lower 'signal-to-noise' ratio in thought-disordered schizophrenic patients [Spitzer *et al.*, 1993]. Such studies bear direct relevance to Cohen and Servan-Schreiber's concept of neuromodulation [1992a], and offer another possible interface between modeling and experimental studies. Finally, much hope for further advancement relies on the rapid development of functional imaging techniques (see [David *et al.*, 1994] for a review of their implications for psychiatry). It should be noted, however, that there is still a significant discrepancy between the scale of the distributed networks of brain activation revealed by current functional imaging studies and the scale of current neural models. The possible role of large-scale cortical networks in performing various cognitive functions has been recently discussed by [Mesulam, 1990].

3. *Modeling additional psychiatric phenomena:* In addition to continuing the study of the disturbances already mentioned in this paper, a few additional issues seem intriguing. For example, the mechanisms of action of Lithium in manic-depressive disorder (see [Manji *et al.*, 1991] for a detailed review), or the therapeutic effect of electroconvulsive treatment (ECT). The latter seems a 'natural' subject for neural modeling studies. ECT is still employed with considerable success in the treatment of major depression patients who do not benefit from anti-depressant pharmacotherapy. However, its application may induce various cognitive deficits, even when applied unilaterally with brief pulse stimulation. As stated by [Khan *et al.*, 1993] in a recent review of cognitive deficits following ECT, studying the induction of cognitive deficits by ECT on a neural level seems currently the most promising approach to this quandary. Finally, as claimed by [Globus & Arpia, 1994], psychiatry is replete with time series data on various levels, such as clinical, physiological, chemical and behavioral. As proposed in [Luciano *et al.*, 1994a], this data may be utilized in an attempt to construct a detailed model of 'the inner workings of the black box', but this seems yet to be a very complex task.

Psychiatric diagnostic entities are syndromes, i.e., conglomerates of symptoms with significant overlap, and many symptoms are manifested in a few disorders. For example, typical schizophrenic symptoms such as delusions (and less frequently, hallucinations) are fairly common in patients with Alzheimer's disease [Cummings *et al.*, 1987]. In addition, neuropathological changes typical of Alzheimer are significantly more common in chronic schizophrenic patients than in an age-matched controls [Prohovnik *et al.*, 1993]. Such findings seem to put forward the notion that neural and synaptic alterations may give rise to specific symptoms, irrespective of the general disease process in the context of which they occur. This notion supports the rational underlying most current psychiatric modeling efforts - that it is possible to study the formation of specific symptoms and tasks in isolation from a significantly more demanding aim - that of studying the pathogenesis of a psychiatric disorder as a whole.

The studies reviewed in this paper all fall in the framework of what one may call 'the central dogma' of neural modeling of brain disorders: Given some specific disease, we gather information on pathological changes occurring in patients having the disease, and construct a model that incorporates this microscopic data and accommodates for macroscopic alterations in the network's behavior that model some clinical aspects of the disease studied. This approach may be viewed as 'top-down' modeling. In contrast, it is intriguing to speculate that in the future significant insight into the pathogenesis of psychiatric disturbances may be gained via an opposite, 'bottom-up' approach. This futuristic scenario hinges upon the development of 'neuromorphic' robots, capable of complex behavior. With such robots in hand, one could alter the workings of the neural networks governing their behavior, and explore the resulting range of behavioral disturbances. Until then, however, there is still plenty to do.

Acknowledgment: I am very grateful to David Horn and James Reggia for collaborating with me in our modeling studies of Alzheimer and schizophrenia. Many thanks are also due to Daniel Amit, Ralph Hoffman, Isaac Meilijson, Manfred Spitzer, Marius Usher and Hezy Yeshurun for many stimulating discussions and helpful comments on these issues.

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] J.A. Anderson. Cognitive and psychological computation with neural models. *IEEE Trans. on Systems,Man, and Cybernetics*, 13 (5):799–815, 1983.

[4] N.C. Andreasen. The mechanisms of schizophrenia. *Current Opinion in Neurobiology*, 4:245–251, 1994.

[5] 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.

[6] A. Avni, E. Ruppin, M. Stern, and H. Munitz. Cognitive function in delirium: a neural model. 1995. Preprint.

[7] D.E. Broadbent. *Decision and stress.* Academic Press, 1971.

[8] E. Callaway and S. Naghdi. An information processing model for schizophrenia. *Arch. of Gen. Psych.*, 39:339–347, 1982.

[9] E. Callaway, R. Halliday, H. Naylor, L. Yano, and K. Herzig. Drugs and human information processing. *Neuropsychopharmacology*, 10(1):9–19, 1994.

[10] E. Callaway. Schizophrenia and interference: an analogy with a malfunctioning computer. *Arch. of Gen. Psych.*, 22:193–208, 1970.

[11] M. Carlsson and A. Carlsson. Interactions between glutamatergic and monoaminergic systems within the basal-ganglia - implications for schizophrenia and Parkinson's disease. *Trends in Neuroscience*, 13:272–276, 1990.

[12] W.T. Carpenter and R.W. Buchanan. Schizophrenia. *New England Journal of Medicine*, 330 (10):681–690, 1994.

[13] J.R. Carrie. Evaluation of a neural network model of amnesia in diffuse cerebral atrophy. *British Journal of Psychiatry*, 163:217–222, 1993.

[14] 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.

[15] J.D. Cohen and D. Servan-Schreiber. Introduction to neural network models in psychiatry. *Psychiatric Annals*, 22(3):113–118, 1992.

[16] J.D. Cohen and D. Servan-Schreiber. A neural network model of disturbances in the processing of context in schizophrenia. *Psychiatric Annals*, 22:131–136, 1992.

[17] J.A. Cohen and D. Servan-Schreiber. A theory of cognitive function and its role in cognitive deficits in schizophrenia. *Schizophrenia Bulletin*, 19(1):85–104, 1993.

[18] J.D. Cohen and D. Servan-Schreiber. Schizophrenic deficits: Neurolopetics and the prefrontal cortex – A reply. *Schizophrenia Bulletin*, 20(3):417–421, 1994.

[19] J.D. Cohen, D. Servan-Schreiber, and J.L. McClelland. A PDP approach to automaticity: principles and models. *American journal of Psychology*, 105:239–269, 1992.

[20] J.D. Cohen, R.D. Romero, D. Servan-Schreiber, and M.J. Farah. Disengaging from the disengage function: the relation of macrostructure to microstructure in parietal attentional deficits. *Journal of Cognitive Neuroscience*, 1994. In press.

[21] J.L. Cummings, B. Miller, M.A. Hill, and R. Neshkes. Neuropsychiatric aspects of Multi-infarct dementia and dementia of the Alzheimer type. *Arch. of Neurol.*, 44:389–393, 1987.

[22] A. David, A. Blamire, and H. Breiter. Functional magnetic resonance imaging: A new technique with implications for psychology and psychiatry. *Brit. Jour. of Psychiatry*, 164:2–7, 1994.

[23] S. Dehaene and J.P. Changeux. A simple model of prefrontal cortex function in delayed response tasks. *Journal of Cognitive Neuroscience*, 1:244–261, 1989.

[24] S. Dehaene and J.P. Changeux. The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, 1:62–79, 1991.

[25] S. T. DeKosky and S.W. Scheff. Synapse loss in frontal cortex biopsies in alzheimer's disease: Correlation with cognitive severity. *Ann. Neurology*, 27(5):457–464, 1990.

[26] M. Farah and J. McClelland. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *J. Exper. Psychol.*, 120:339–357, 1991.

[27] M.J. Farah and J.L. McClelland. Neural network models and cognitive neuropsychology. *Psychiatric Annals*, 22(3):148–153, 1992.

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

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

[30] B. Gordon. Confrontation naming: Computational model and disconnection simulation. In M. Arbib, D. Kaplan, and J. Marshall, editors, *Neural models of language processes*, pages 511–530. Academic Press, 1982.

[31] 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.

[32] S. Grossberg. Some psychological correlates of a developmental, cognitive and motivational theory. *Annals of the new york academy of sciences*, 425:58–142, 1984.

[33] L. A. Hansen, R. DeTeresa, P. Davies, and R.D. Terry. Neocortical morphometry, lesion counts, and choline acetyltransferase levels in the age spectrum of alzheimer's disease. *Neurology*, 38:48–54, 1988.

[34] M.E. Hasselmo and E. Barkai. Cholinergic modulation of the input/output function of rat pyriform cortex pyramidal cells. *Soc. Neurosci. Abstr.*, 18:521, 1992.

[35] M.E. Hasselmo and J.M Bower. Cholinergic suppression specific to intrinsic not afferent fiber synapses in the rat piriform (olfactory) cortex. *Journal of Neurophsiology*, 67(5):1222–1229, 1992.

[36] M.E. Hasselmo, B.P. Anderson, and J.M. Bower. Cholinergic modulation of cortical associative memory function. *Journal of Neurophsiology*, 67(5):1230–1246, 1992.

[37] M.E. Hasselmo. Acetylcholine and learning in a cortical associative memory. *Neural computation*, 5(1):32, 1993.

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

[39] D. Hestenes. A neural network theory of manic-depressive illness. In D.S. Levine and S.J. Leven, editors, *Motivation, Emotion and Goal Direction in neural networks.* Lawrence Elbaum Associates, Publishers, 1992.

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

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

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

[43] R.E. Hoffman, J. Rapaport, R. Ameli, T. H. McGlashan, D. Harcherik, and D. Servan-Schreiber. The pathophysiology of hallucinated voices and associated speech perception impairments in schizophrenic patients. 1994. Preprint.

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

[45] R.E. Hoffman. Attractor neural networks and psychotic disorders. *Psychiatric Annals*, 22(3):119–124, 1992.

[46] D. Horn and E. Ruppin. Extra-pyramidal symptoms in alzheimer's disease: a hypothesis. *Medical Hypothesis*, 39:316–318, 1992.

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

[48] 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.

[49] D. Horn, N. Levy, and E. Ruppin. Local mechanisms of synaptic compensation and deletion in Alzheimer's disease. 1995. Preprint.

[50] G.W. Humphreys, T. Freeman, and H. Muller. Lesioning a connectionist model of visual search. *Canadian Journal of Psychology*, 46:417–460, 1992.

[51] T.H. Jobe, M. Harrow, E.M. Martin, H.J. Whitfield, and J. R. Sands. Schizophrenic deficits: Neurolopetics and the prefrontal cortex. *Schizophrenia Bulletin*, 20(3):413–416, 1994.

[52] M.H. Joseph, C.D. Frith, and J.L. Waddington. Dopaminergic mechanisms and cognitive deficits in schizophrenia. *Psychopharmacology*, 63:273–280, 1979.

[53] R. Katzman. Alzheimer's disease. *New England Journal of Medicine*, 314(15):964–973, 1986.

[54] A. Khan, M.H. Mirolo, H.A. Mirolo, and S. Miller. Can ECT-induced cognitive effects be altered pharmacologically? *Prog. Neuro-Psychophrmacol. & Biol. Psychiat.*, 17:861–873, 1993.

[55] S.J. Leven. Learned helplessness, memory, and the dynamics of hope. In D.S. Levine and S.J. Leven, editors, *Motivation, Emotion and Goal Direction in neural networks.* Lawrence Elbaum Associates, Publishers, 1992.

[56] D.S. Levine and P.S. Prueitt. Modeling some effects of frontal lobe damage – novelty and preseveration. *Neural Networks*, 2:103–116, 1989.

[57] D. Li and D. Spiegel. A neural network model of dissociative disorders. *Psychiatric Annals*, 22(3):144–147, 1992.

[58] J.S. Luciano, M.A. Cohen, and J.A. Samson. A neural model of unipolar depression. In *Irish Neural Network Conference*, September 12-13 1994.

[59] J.S. Luciano, M.A. Cohen, J.A. Samson, and P.G. Hagan. A prototype and non linear neural network approach to subtyping depressive illness by initial input data and treatment outcomes. In *Irish Neural Network Conference*, September 12-13 1994.

[60] A.J. Mandell and K.A. Selz. Dynamical systems in psychaitry: Now what? *Biological Psychiatry*, 32:299–301, 1992.

[61] H.K. Manji, J.K. Hsiao, E.D. Risby, J. Oliver, M.V. Rudorfer, and W.Z. Potter. The mechanisms of action of Lithium. *Arch. of Gen. Psych.*, 48:505–512, 1991.

[62] N. Martin, G. Dell, E.M. Safran, and M.F. Schwartz. Origins of paraphasias in deep dysphasia: Testing the consequences of decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language*. To appear.

[63] J.L. McClelland and D.E. Rumelhart. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT press, 1986.

[64] J.L. McClelland and D.E. Rumelhart. Amnesia and distributed memory. In J.L. McClelland and D.E. Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2*, page 170. MIT press, 1986.

[65] J.L. McClelland, B.L. McNaughton, and R.C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from successes and failures of connectionist models of learning and memory. *The Journal of Neuroscience*, 1994. Submitted.

[66] J.L. McClelland. Connectionist models and psychological evidence. *Journal of Memory and Language*, 27:107–123, 1988.

[67] M. Mehta, C. Dasgupta, and G. Ulal. A neural network model for kindling of focal epilepsy: basic mechanism. *Biol. Cybern*, 68:335–340, 1993.

[68] M-Marsel Mesulam. Large scale neurocognitive networks and distributed processing for attention, language and memory. *Annals of Neurology*, 28:597–613, 1990.

[69] R.C. O'Reilly and J.L. McClelland. Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *preprint*, 1994.

[70] K.E. Patterson, M.S. Seidenberg, and J.L. McClelland. Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R.G.M Morris (Eds.), editor, *Parallel distributing processes: implications for psychology and neuroscience*, pages 131–181. Oxfrod University Press, 1990.

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

[72] J.W. Pettegrew, M.S. Keshavan, K. Panchalingam, S. Strychor, D.B. Kaplan, M.G. Tretta, and M. Allen. Alterations in brain high-energy phosphate and membrane phospholipid metabolism in first-episode, drug-naive schizophrenics. *Arch. of Gen. Psych.*, 48:563–568, 1991.

[73] E.M. Pich and P.D. Guidice. Can neural networks be used as models of neuropsychological dysfunctions? *Int. Jour. of Neural Systems*, 3 (Supp. 1992):163–168, 1992.

[74] P. Pichot. Nosological models in psychiatry. *Br. Jour. of Psychiatry*, 164:232–240, 1994.

[75] D.C. Plaut and T. Shallice. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive neuropsychology*, 10(5):377–500, 1993.

[76] D.C. Plaut and T. Shallice. Persevarative and semantic influences on visial object naming erors in optic aphasia: A connectionist account. *Journal of cognitive neuroscience*, 5(1):89–117, 1993.

[77] D.L. Price, D.R. Borchelt, and S.S. Sisodia. Alzheimer disease and the prion disorders amyloid beta-protein and prion protein amyloidoses. *Proc. Natl. Acad. Sci.*, 90(14):6381–6384, 1993.

[78] I. Prohovnik, A.J. Dwork, M.A. Kaufman, and N. Wilson. Alzheimertype neuropathology in elderly schizophrenia patients. *Schizophrenia Bulletin*, 19(4):805–816, 1993.

[79] J. Reggia and R. Berndt. Modelling reading aloud and its rlevance to acquired dyslexia. *Computers methods and rpograms in biomedicine*, 22:13–19, 1986.

[80] J. Reggia, R. Berndt, and L. D'Autrechy. Connectionist models in neuropsychology. In *Handbook of Neuropsychology*, volume (9), pages 297–333. Elsevier Science, Amsterdam, 1994.

[81] G.W. Roberts. Schizophrenia: the cellular biology of a functional psychosis. *Trends in Neuroscience*, 13 (6):207–211, 1990.

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

[83] E. Ruppin, J. Reggia, and D. Horn. A neural model of positive schizophrenic symptoms. *Schizophrenia Bulletin*, 1995. To appear.

[84] D. Servan-Schreiber and J. Cohen. A neural network model of catecholamine modulation of behavior. *Psychiatric Annals*, 22(3):125–130, 1992.

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

[86] David Servan-Schreiber. From physiology to behavior: Computational models of catecholamine modulation of information processing (ph.d. thesis). Technical Report TR CMU-CS-90-167, School of Computer Science, Carnegie Mellon University, 1990.

[87] J. Shrager, T. Hogg, and B.A. Huberman. Observation of phase transitions in spreading activation networks. *Science*, 236:1092–1094, 1987.

[88] M. Spitzer, P. Bohler, M. Weisbrod, and U. Kischka. A neural network model of phantom limbs. *Biological Cybernetics*. To appear.

[89] M. Spitzer, U. Braun, L. Hrmle, and S. Maier. Associative semantic network dysfunction in thought-disordered schizophrenic patients: Direct evidence from indirect semantic priming. *Biological Psychiatry*, 34:864–877, 1993.

[90] J.R. Stevens. An anatomy of schizophrenia? *Arch. of Gen. Psych.*, 29:177–189, 1973.

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

[92] G. Sutton, J. Reggia, S. Armentrout, and C. D'Autrechy. Map reorganization as a competitive process. *Neural Computation*, 6:1–13, 1994.

[93] S.L. Thaler. Virtual input phenomena within the death of a simple pattern associator. *Neural Networks*, 8(1):55–65, 1995.

[94] 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.

[95] M. Tsukuda. Theoretical model of the hippocampal-cortical memory system motivated by phsiological functions in the hippocampus. *Cybernetics and systems*, 25:189–206, 1994.

[96] S. Vinogradov, R.J. King, and B.A. Huberman. An associatonist model of the paranoid process: application of phase transitions in spreading activation networks. *Psychiatry*, 55:79–94, 1992.

[97] M.A. Virasoro. The effect of synapses destruction on categorization by neural networks. *Europhys. Lett.*, 7:293–298, 1988.

[98] J.L. Waddington. Neurodynamics of abnormalities in cerebral metabolism and structure in schizophrenia. *Schizophrenia Bulletin*, 19(1):55–69, 1993.

[99] L. Wang and J. Ross. *Proc. Natl. Acad. Sci USA*, 87:7110, 1990.

[100] L. Wang and J. Ross. *Biol. Cybern.*, 64:231, 1991.

[101] D.R. Weinberger. Implications of normal brain development for the pathogenesis of schizophrenia. *Arch. of Gen. Psych.*, 44:660–669, 1987.

[102] R. Wong, R. Traub, and R. Miles. Cellular basis of neuronal synchrony. *Advances in Neurology*, 44:583–592, 1986.

[103] C.C. Wood. Variations on a theme by Lashley: lesion experiments on the neural model of Anderson, Silverstien, Ritz and Jones. *Psychological Review*, 85(6):582–591, 1978.