Part one

General concepts

1

Neural networks and psychopathology: an introduction

DAN J. STEIN and JACQUES LUDIK

The recent shift in psychiatry from a predominantly psychodynamic model towards a neurobiological paradigm has led to important advances in our understanding and management of many mental disorders. At the same time, this shift has been characterized as a move from a brainless psychiatry to a mindless one (Lipowski, 1989). Certainly, the continued existence of different psychiatric schools with widely divergent approaches to psychopathology and its treatment suggests that psychiatry continues to lack an adequate theoretical underpinning.

During the same time that psychiatry has undergone a paradigm shift, academic psychology has also experienced a revolution – the so-called cognitive revolution against behaviorism (Gardner, 1985). Cognitive science, a multidisciplinary arena encompassing cognitive psychology, artificial intelligence, neuroscience, linguistics, anthropology, and philosophy, and based on computational models of the mind, is now a predominant approach. Not surprisingly, clinicians have asked whether the constructs and methods of cognitive science are also applicable to psychopathology.

Indeed, a promising dialogue between clinical and cognitive science has emerged (Stein and Young, 1992). Both cognitive–behavioral therapists and psychodynamic researchers have increasingly drawn on cognitivist work in their theoretical and empirical studies of psychopathology and psychotherapy. Schema theory, for example, has been applied to a range of clinical phenomena (Stein, 1992). Such cognitivist work is often immediately attractive to the clinician insofar as it incorporates a range of theoretical disciplines and insofar as it is based on hard empirical studies.

One of the most important developments in modern cognitive science has been connectionism, the field concerned with neural network models (Rumelhart et al., 1986a). Whereas early work in cognitive science emphasized 'top-down' symbolic architectures and the manipulation of mental representations, connectionism has focused on 'bottom-up' models that specify the interactions of simple processing units. In contrast to the serial processing of traditional symbolic models, in neural networks information processing occurs simultaneously in all units (parallel distributed processing). Increasingly, neural networks are being applied in the clinical arena, again offering the clinician a set of constructs and methods that seem sophisticated and robust (Hoffman, 1987; Park and Young, 1994).

This book provides a forum for the presentation of pioneering work at the intersection of clinical science and connectionism. This introductory chapter details some of the defining features of the connectionist paradigm, and considers some of the advantages and possible limitations of this approach for clinical science.

Features of neural networks

Connectionist models focus on sets of processing units (idealized neurons) and their interactions. Some of the earliest connectionist work was done by Donald Hebb (1949) in his speculations about the basis of neuronal functioning. He put forward the idea of cell assemblies, and proposed that simultaneous activation of two cells resulted in strengthening of their connection (Hebb's rule). Other theorists helped develop sophisticated mathematical theories to describe such neuronal networks (Grossberg, 1980; McCulloch and Pitts, 1943; Rosenblatt, 1962; Selfridge and Neisser, 1960), and the development of modern computers allowed ready implementation of detailed connectionist models.

Any particular neural network model can be described in terms of its specific processing units, the way these are put together, and the way in which they learn (Hanson and Burr, 1990). Like neurons, each unit has inputs (dendrites) from other units, and outputs (axons) to other units. Each input has a particular weight (synapse), which can be positive (excitatory) or negative (inhibitory). Whether or not a unit is activated is determined by this net input and by its current activation.

The topology of a neural network is the way in which units are joined to one another. In a totally connected network, such as the Hopfield network (Hopfield, 1982, 1984), all units are connected to one another (Fig. 1.1). In a feedforward unit, information flows in only one direction, from input units to output units. In multilayer networks, there are also hidden units between input and output units (Fig. 1.2).

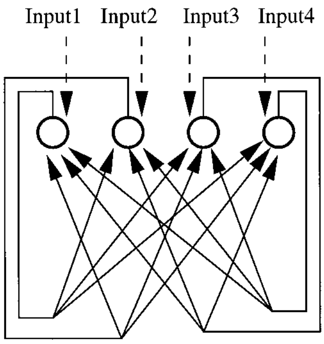

Fig. 1.1 A Hopfield neural network in which all units are connected to one another.

Learning takes place in networks via modification of synaptic weights. Neural networks, for example, can be trained to associate particular input patterns with particular output patterns. During training, input patterns are presented and synaptic weights are changed according to a learning rule. In a multilayer network, error can be measured across the output units and then compensatory changes can be made at each level of the network (back-propagation).

How are memories stored in a network? Many networks can be conceptualized as constraint networks in which each unit represents a hypothesis (i.e., a feature of the input), and in which each connection represents constraints among the hypotheses (Rumelhart et al., 1986b). A variation of Hebb's rule, for example, states that if features A and B often co-exist, then the connection between the two will be positive. On the other hand, when the two features exclude one another, then the connection will be negative. When the network runs, it settles into a locally optimal state in which as many as possible of the constraints are satisfied.

The information processing of a network from state to state can be conceptualized in terms of movement over a goodness-of-fit landscape (Rumelhart et al., 1986b). The system processes input by shifting from state to state until it reaches a state of maximal constraint satisfaction, that is, it climbs upward until it reaches a goodness maximum. A landscape can be described in terms of the set of maxima which the system can find, the size of the region that feeds into each maximum, and the height

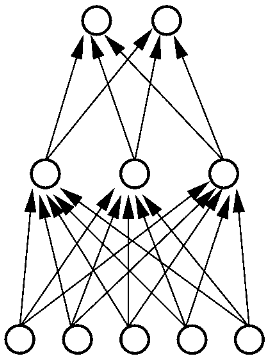

Fig. 1.2 A multilayer neural network in which hidden units intervene between input and output unit.

of the maximum itself (Fig. 1.3). The positions of the system correspond to the possible interpretations, the peaks in the space correspond to the best interpretations, the extent of the foothills surrounding a particular peak determines the likelihood of finding the peak, and the height of the peak corresponds to the degree that the constraints of the network are actually met (Rumelhart et al., 1986b).

Schemas versus neural networks

Characterizing neural networks in terms of a goodness-of-fit landscape has immediate intuitive appeal. This characterization allows neural networks to be compared with schemas – cognitivist constructs that are, as noted earlier, increasingly familiar to clinicians. A schema is a prototypical abstraction that develops from past experience and that guides the organization of new information (Thorndyke and Hayes-Roth, 1979; Stein, 1992). Schemas allow rapid processing of information, but also result in typical biases (Winfrey and Goldfried, 1986).

Similarly, a particular neural network, prompted by a given set of data, rapidly moves toward a previously acquired landscape. This allows rapid information processing, but, again, may result in certain distortions (Rumelhart et al., 1986b). This view of schemas is perhaps more fluid than the conventional one; for example, schemas can be defined as inflexible (narrow peaks in the goodness-of-fit landscape) or more flexible (with

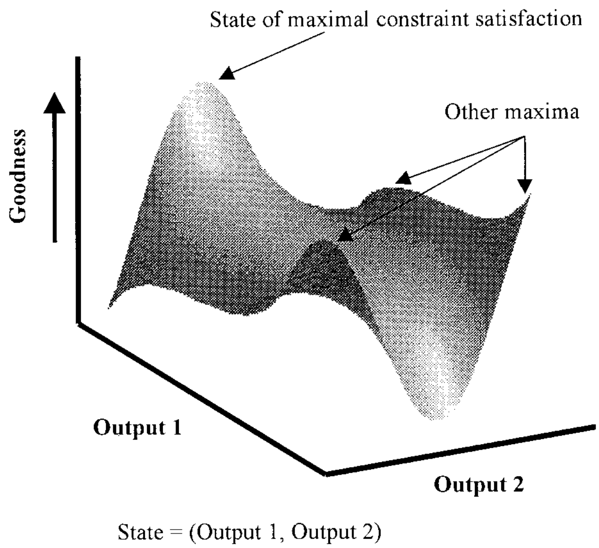

Fig. 1.3 The goodness-of-fit landscape for a neural network as function of the activation of two output units. A state is represented here by an (output 1, output 2) pair. For every state of the network, the pattern of inputs and connection weights determines a value of the goodness-of-fit function. The network processes its input by moving upward from state to state until it reaches a state of maximum goodness.

broad plateaus allowing for movement in the region of the maximum) (Rumelhart et al., 1986b).

Consider, for example, a woman who has been abused in childhood. She may develop a mistrust schema according to which others are not easily to be trusted. She is consequently liable to bias her interpretation of reality in particular ways, perhaps drawing false generalizations about authority figures or viewing neutral situations as unsafe. Both schemas and neural networks provide a way of explaining how such biases are 'built in', without having to rely on explicit cognitive rules.

Rumelhart et al. (1986b) conclude that the relationship between neural network models and schema models is largely a matter of a different degree of analysis. Whereas schema models are predominantly 'top-down' in their approach, neural network models work from the 'bottom-up'. An advantage of the neural network approach is its ability to demonstrate in fine detail how cognitive phenomena emerge from interactions of simple elements of the system.

Similarly, in the clinical situation, the neural network approach may allow a better understanding of the microstructure of schemas. While schemas have allowed an integration of different kinds of theories, they have been less successful at incorporating neurobiological information than some would have hoped. For example, while a theory of mistrust schemas does exist (Young, 1990), this is not easily able to incorporate such clinical data as the efficacy of psychotropic medication in patients with personality disorders. This kind of data might be better understood if the cognitivist model used was a bottom-up one.

The grounding problem

Why should neural network modelers pay attention to clinical science? It may be argued that clinical phenomena are particularly challenging insofar as they necessarily encompass a broad range of levels of analysis. Thus, it may be argued that such phenomena demand a level of analysis that extracts the maximum potential from cognitive science.

There remains, for example, a basic problem in cognitive science that may be characterized (following Harnad, 1990) as the symbol-grounding problem. This concerns how the meanings of the symbols in a system can be grounded so that they have meaning independently of an external interpreter. This problem may lie at the heart of a number of important debates in cognitive science.

Consider, for example, Searle's (1980) well-known argument against symbolic models of the mind. Searle notes that while a Turing computer could conceivably implement a range of rules necessary for translating Chinese symbols into English ones, it cannot be argued that this computer understands Chinese. For example, while it might be possible for a person to memorize all the syntactical rules employed by such a program, this would not necessarily mean that the person had a grasp of the semantic meanings of Chinese.

Similarly, a range of so-called situated cognitivists argue against conventional symbolic cognition (Norman, 1993). Symbolic cognitivists hold that cognitive processing essentially involves the manipulation of symbols, and that the task of cognitive science is to provide formalized descriptions of these transformations. Situational cognitivists hold that cognitive processes necessarily take place within a particular interactive social context, and that the task of cognitive science is to understand how cognitive processes are situated in experience.

A clinical example may be useful here. Consider once again the mistrustful patient. In a pioneering project in early artificial intelligence, Colby (1981) developed a computer program, PARRY, which simulated paranoid thought processes. The program relied on a symbolic architecture, and incorporated a number of rules that governed the manipulation of symbols. For example, one rule stated that when self-esteem score decreased, level of suspicion would be increased. PARRY was highly successful insofar as experienced clinicians interacting with it were sometimes unable to tell whether they were dealing with a computer or with a person (the Turing Test).

From Searle's perspective, however, the claim that PARRY is in fact paranoid should not be taken seriously. PARRY is a computer program that implements syntactic rules, but that lacks semantic understanding and intentionality. Similarly, from the perspective of a situated cognitivist, although it might be conceded that PARRY devotes attention to the interpersonal context of paranoid behavior, its focus on the manipulation of symbols means that it ultimately fails to come to terms fully with this phenomenon.

Indeed, from a clinical viewpoint, although PARRY was a pioneering contribution to the intersection between clinical and cognitive science, and although it provided an interesting hypothesis for and test of the cognitive processes underlying paranoia, its success was only partial. In particular, PARRY ignored many aspects of the clinical phenomenon of paranoid thinking, including data on the neurobiology of psychosis. Ultimately, PARRY was unable to explain the underlying mechanisms upon which its rules were based.

Harnad (1989) has proposed a variant of the Turing Test, the Total Turing Test (TTT), in order to help solve the symbol-grounding problem. In addition to simulation of pen-pal (symbolic) interactions, passing the TTT demands simulation of 'robotic capacity' – all of our sensorimotor capacity to discriminate, recognize, identify, manipulate, and describe the world we live in. Harnad argues that a system is grounded only when it has the symbolic and robotic capacity to pass the TTT in a coherent way, that is when its symbolic capacity is grounded in its robotic capacity rather than being mediated by an outside interpretation projected onto it.

To return to the Chinese room, the question is no longer whether the Turing Test candidate really understands Chinese or can merely be interpreted as if he or she were understanding it, but rather whether the TTT candidate (a robot with optical transducers) really sees the Chinese letters and really writes the English version down or whether it is merely interpretable as if it were doing this. If Searle now attempted to implement the

TTT candidate without seeing or writing (as he attempted to implement the TT candidate without understanding), this would be impossible.

Insofar as real transduction is essential to TTT capacity, a TTT candidate would satisfy the demands of a situated cognitivist. This candidate would not simply be manipulating symbols, but would in fact demonstrate cognitive processes that were situated in experience. Such cognitive processes could no longer be said to be independent of their physical instantiations (as symbolic cognitivists are so fond of averring).

Similarly, from a clinical viewpoint, modeling sensory transduction is indeed necessary if psychopathological phenomena such as paranoia are to be fully understood. Cognitive clinical science needs to take cognizance of the growing emphasis by cognitivists on the embodiment of cognition (Lakoff, 1987), both in the sense of being embodied within the physicality of the brain and in the sense of being embodied within particular social situations. Given the increased understanding of the neurobiology underlying psychopathology, models that incorporate this kind of understanding may well be possible.

Consider, for instance, a pioneering example of work at the intersection of connectionism and the clinic, the research of Jonathan Cohen and David Servan-Schreiber (1992) on schizophrenia. In their model of this disorder they model how changes in neurotransmitter function (dopamine gain) result in dysfunction on neuropsychological testing. The model therefore moves toward providing a seamless integration of the neurobiology and psychology of this complex disorder. Although the model only attempts to cover limited aspects of schizophrenia and could not pass the TTT, it does not simply involve syntactic rule transformation, and it provides a preliminary account of how psychopathological processes in schizophrenia are in fact embodied.

Some difficulties

Different kinds of neural network models may be applicable to different arenas within psychiatry. For example, there is currently work on neural network approaches to diagnosis (see Chapter 3), neural network modeling of psychopharmacological data (see Chapter 4), and neural network modeling of psychotherapeutic processes (See Chapter 5) and psychodynamic phenomena (see Chapter 10). Nevertheless, much work follows the pioneering lines taken by Hoffman (1987) and Cohen and Servan-Schreiber (1992), in which 'lesions' to neural network models are made in the hope of simulating psychopathological data.

It may be argued that there remains an important disjunction between phenomena as they are witnessed by the clinician (e.g., schizophrenic delusions) and the kind of inputs and outputs that are processed by neural networks (e.g., numerical patterns, results on a neuropsychological test). Psychiatric research methodologies such as functional brain imaging are currently allowing insights into the concrete mechanisms that underlie specific clinical phenomena (e.g., basal ganglia activation in obsessions and compulsions); in contrast, neural networks are experience distant.

This distance between clinical experience and neural network analysis may also account for the worrying fact that particular neural networks are used by various authors to account for a range of different psychopathological phenomena. For example, Cohen and Servan-Schreiber's network for schizophrenia has also been used to account for other kinds of phenomena including obsessive–compulsive disorder (OCD). Certainly, a single neurotransmitter system may in fact be involved in several different psychiatric disorders. However, the use of a single neural network to explain diverse clinical phenomena also appears to suggest that it can explain everything.

To some extent, however, a similar issue arises in schema theory. Is there anything about minds that schema theory does not explain? Given that schemas and neural networks seem to incorporate general rules of cognition, their application to any clinical phenomena will perhaps always result in at least a partial ring of truth. The trick for future researchers will be to specify the details of these applications with increasing depth, so that specific differences in the neural networks/schemas of different clinical phenomena become increasingly clear.

So much for the issue of phenomenology. What are the objections that a strict neurobiological approach may have for neural network theory? It seems clear that many processes in the brain do operate according to the principles of parallel distributed processing. Nevertheless, neural network models of clinical phenomena typically fail to incorporate many of the fine details of neurobiological knowledge, and they may even directly contradict the findings of modern neuroscience. For example, the fact that units typically have both inhibitory and excitatory connections is at odds with neurobiological data that most neurons are either inhibitory or excitatory.

However, this criticism fails to take adequate account of the level of analysis that neural network models hope to achieve. While computational models of neurophysiological and neuropathological processes are often extremely relevant to psychiatry, neural networks that aim to model