Place, U. T. (1993). *Following 'the natural lines of fracture': Concept formation in neural networks* [Conference presentation, presented at the Symposium on Associationism, Behaviour Analysis and Connectionism, held at the Annual Conference of the Experimental Analysis of Behaviour Group, University College, London 31st March 1993].

FOLLOWING "THE NATURAL LINES OF FRACTURE": CONCEPT FORMATION IN NEURAL NETWORKS[1]

Ullin T. Place

University of Wales Bangor

Abstract

It is an implication of Darwin's theory of evolution by variation and natural selection that the survival and reproduction of complex free-moving living organisms, animals in other words, depends on their ability to change the spatial relations between themselves and other objects, including other organisms of the same and of different species, and so bring about the conditions necessary for that survival and reproduction. In order to do that the organism requires a system - its nervous system - whose function is to match the output to the current stimulus input on the one hand and the organism's current state of deprivation with respect to conditions required for its survival and successful reproduction on the other. Matching behaviour to the conditions required for survival and reproduction is the function of the motivational/emotional part of the system. Matching behaviour to current stimulus input is the function of the sensory/cognitive part of the system. The sensory/cognitive system cannot perform its function successfully without the ability to group inputs together in such a way that every actual and possible member of the class or category so formed is a reliable indicator of the presence of an environmental situation in which a particular behavioural strategy or set of such strategies is going to succeed. In other words the survival and reproduction of an organism of this kind depends crucially on its having a conceptual scheme, a conceptual scheme moreover, which reliably predicts the actual behaviour-consequence relations operating in the organism's environment.

Although verbs such as 'classifying', 'categorizing' and 'conceptualizing' are not to be found in Skinner's writings, there is an important passage in *The Behavior of Organisms* (Skinner 1938) where he addresses the issue which others talk about when they use such terms. Thus in Chapter One, after outlining his "System of Behavior", he goes on to say

> The preceding system is based upon the assumption that both behavior and environment may be broken into parts which retain their identity throughout an experiment and undergo orderly changes. If this assumption were not in some sense justified, a science of behavior would be impossible. But the analysis of behavior is not an act of arbitrary sub-dividing. We cannot define the concepts of stimulus and response quite as simply as 'parts of behavior and environment' without taking account of the natural lines of fracture along which behavior and environment actually break. (Skinner 1938 p.33).

What Skinner has primarily in mind in this passage is the way the scientist's concepts need to be shaped into conformity with what he calls "the natural lines of fracture." But on the Darwinian argument the same must be true of the stimulus classes within which any living organism's behaviour generalises and between which it discriminates. It is argued that studying the properties of artificially constructed neural networks helps us to understand how the brain develops patterns of generalisation and discrimination which do indeed "follow the natural lines of fracture along which behavior and environment actually break." Attention is drawn to the role of the 'hidden layer' in responding to resemblances of pattern, to the role of re-entrant/recurrent and reverberatory circuits in establishing

[1] Presented at a symposium on 'Associationism, Behaviour Analysis, and Connectionism' at the Annual Conference of the Experimental Analysis of Behaviour Group, University College, London, 31st March 1993.

expectations on the basis of consecutive stimulus patterns, and to the role of error-correction in bringing stimulus classes into line with the contingencies experienced during learning.

Stimulus classes and response classes as intensional sets

This paper picks up a point which I made towards the end of the paper which I presented this morning.[2] Those of you who attended that presentation will recall that among the forms of mediating behaviour which I suggested, need to be postulated in order to account for the problem-solving abilities of pre-linguistic organisms was a process of categorisation or concept formation whereby the stimulus events impinging on the organism's receptors are organised into what Skinner (1938) calls "stimulus classes". A stimulus class is what I call (following a suggestion from Professor Jay Moore of the University of Wisconsin, Milwaukee - personal communication), 'an intensional set'. An *intensional set* contrasts with an *extensional class* in that it includes, as the extensional class does not, a range of possible future instances, as well as actual instances which have occurred in the past or are occurring as we speak. In other words, a stimulus class is a *disposition* on the part of the organism to respond in much the same way in the presence of any stimulus which displays the property or set of properties which define the class and not to respond in that way in the presence of otherwise similar stimuli which lack that property or set of properties.

By the same token, a *response class* is an intensional set consisting of a range of possible future behaviour patterns which are topographically similar and whose occurrence is potentiated in the presence of an instance of the same stimulus class or set of stimulus classes.

Another way of putting the same point is to say that a stimulus class is the range of stimuli across which a particular response class or set of response classes *generalizes* and beyond which the organism *discriminates* between one stimulus class and another.

The 'natural lines of fracture'

It is in the context of his introduction of the concepts of 'stimulus class' and 'response class' that Skinner makes the remark which I quoted this morning and which gives this paper its title.

[2] ["Is there an operant analysis of animal problem-solving". See also https://utplace.uk/search-results/?search_field=anchor&value=place-1992b]

> We cannot define the concepts of stimulus and response quite as simply as 'parts of behavior and environment' without taking account of the natural lines of fracture along which behavior and environment actually break. (Skinner 1938 p.33).

As I pointed out this morning, in this passage

> Skinner is talking primarily about the way in which the scientist who is studying behaviour must ensure that his or her conceptual scheme follows 'the natural lines of fracture along which behaviour and environment actually break'.

However, in the light of what John Donahoe calls the "selectionism" which becomes more conspicuous in Skinner's later writings, particularly 'Selection by Consequences' (Skinner 1981), than it is in *The Behavior of Organisms*, it is entirely within the spirit, if not the letter, of Skinner's thought to point out that it is not just the behavioral scientist who must respect 'the natural fracture'. So too, if it is to survive and reproduce itself, must the behaving organism.

Learning to follow the natural lines of fracture

As I again pointed out this morning, the implication of this extension of Skinner's thought from the scientist to the behaving organism is that

> in order to survive and reproduce itself an organism needs a conceptual scheme in the sense of a set of stimulus classes controlling its behaviour which correspond to those resemblances and differences between things in its environment which mark off one set of biologically significant contingencies from another.

As I remarked this morning,

> To some extent, no doubt, the processes of variation and natural selection which Skinner (1975) refers to as "the contingencies of survival" have ensured that organisms are endowed with such a conceptual scheme by their genetic constitution.

But, if the individual organism is to adapt to contingencies other than those which have contributed to the evolution of the species to which it belongs, it must be able to modify and extend its innate conceptual scheme by the process of learning.

Three features of networks which give realism to a conceptual scheme

In this paper, I want to draw attention to three features which are to be found in *some* of the neural networks which have been used by connectionists to model the learning process. Although, as we shall see, the picture is not yet complete, they are all features which give to the networks, including the brain, which possess them,

the ability, given appropriate training, to develop a conceptual scheme made up of stimulus classes which follow the "natural lines of fracture", the pattern of contingencies which actually operate in that environment. These three features are

(1) the layer of *hidden units* or nodes intermediate between the layer of input nodes and the layer of output nodes,

(2) the so-called "*recurrent*" (Jordan 1986) or "*re-entrant*" (Edelman 1987) circuits which feed from the output layer back into the input layer of the same network and thence onto the same hidden layer units as those controlled directly by environmental stimuli,

(3) though I have doubts about the principle of back-propagation as a way of effecting this, the capacity of networks to learn by the process of *trial and error-correction*, sometimes referred to as "supervised learning".

The role of the hidden layer in pattern recognition learning

There is an ancient tradition within psychology which can be traced back through Hull's (1943) account of "primary stimulus generalization" to Pavlov's (1927) introduction of that term and before him through Wundt back to the psychophysics of Fechner (1860) in which stimulus generalization is conceived as proceeding along simple physical dimensions of the energy impinging on the sense organ. It should be obvious that if generalization is to follow the natural lines of fracture, it cannot proceed in this way. The properties of a stimulus to which the organism needs to respond in order to group together those features which indicate the presence or availability of the same contingency are invariably *patterns*, patterns which occur in a wide variety of contexts and observation conditions. It is one of the most striking properties of a neural network that it is capable of generalizing and discriminating and of learning to generalize and discriminate on the basis of resemblances and differences of pattern in a variety of different contexts.

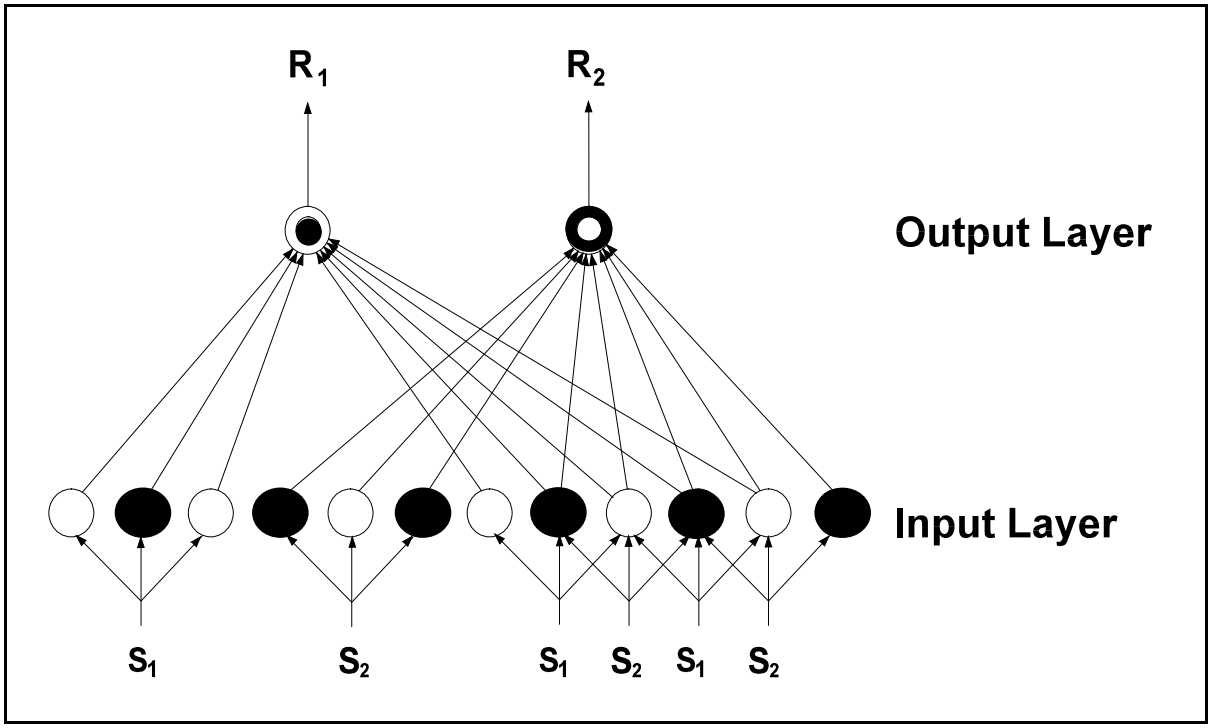

FIGURE 1. TWO LAYER NETWORK WITH STIMULUS GENERALIZATION ONLY

How this is achieved is illustrated on Figure 1. This shows a two layer network in which every other node in the input layer is differentiated from its neighbour by some difference, say, the frequency at which it fires. This difference between alternate input nodes is represented on Figure 1 by colouring them alternately black and white. Its effect is to yield two different stimulus patterns each consisting in the firing of three adjacent nodes.

(a) the property defining stimulus class S_1 is a pattern consisting of two black nodes on the outside with one white node on the inside.

(b) the property defining stimulus class S_2 is a pattern consisting of two white nodes on the outside and one black node on the inside.

The diagram shows overlapping and non-overlapping examples of these two stimulus classes, together with the effective connections between these instances of the two node patterns and the two nodes in the output layer which each pattern triggers.

Figure 1 assumes that the responses triggered by members of these two stimulus classes are two simple alternatives which don't admit of any variation apart from the difference between them.

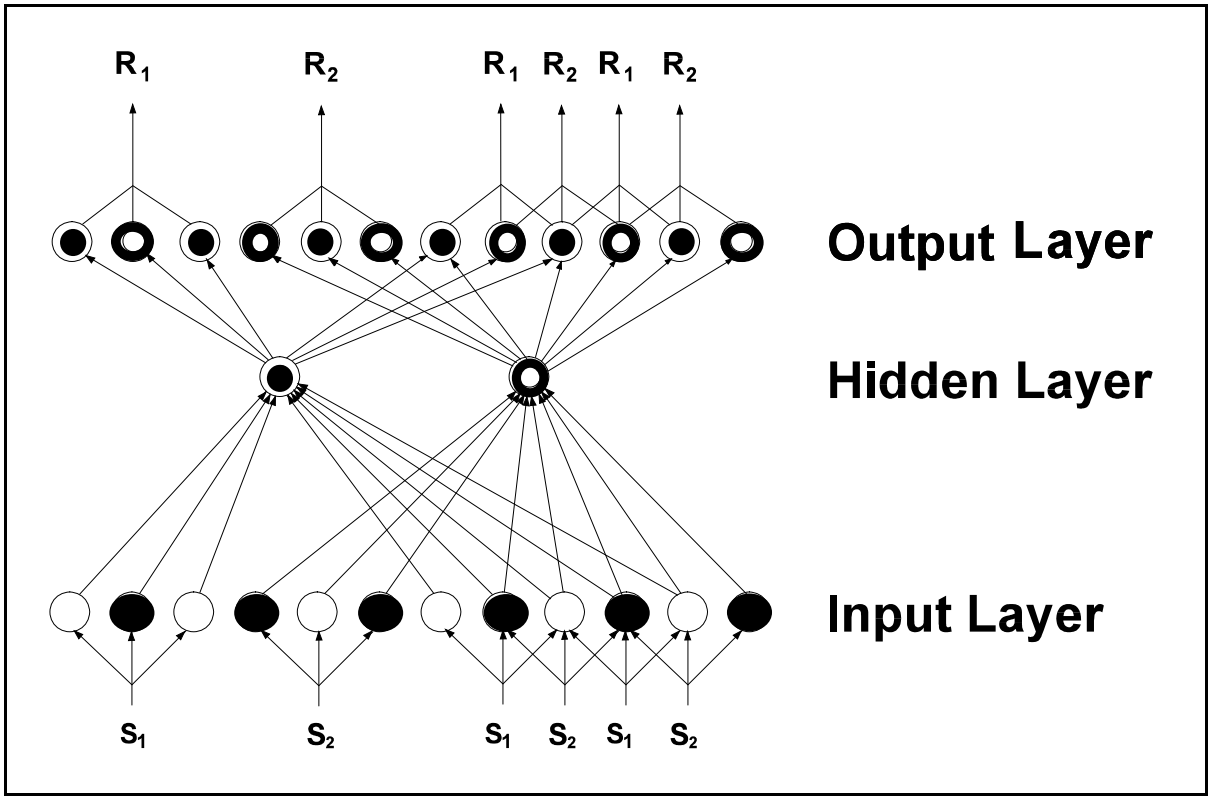

FIGURE 2. THREE LAYER NETWORK WITH STIMULUS *AND* RESPONSE GENERALIZATION

In practice, as Skinner's concept of the response class implies, responses are subject to as much variation and generalisation from context to context as are stimuli. In terms of a network, this situation is represented in Figure 2 which is generated from Figure 1 by converting the former output layer into a hidden or intermediate layer. A new multi-node output layer is added in which various versions of two similar patterns of nodes are triggered by the nodes in what is now the hidden layer. These different versions constitute the response classes R_1 and R_2. The choice between the different versions of the same response is determined by other factors in the prevailing circumstances. In the language of the connectionist, the input-output transformation in the case of Figure 1 is linear; in Figure 2 it is non-linear.

Recurrent/re-entrant circuits and the concept of expectation

As we ordinarily construe the matter, if the onset of a stimulus pattern S_1 is regularly followed within a brief space of time by another stimulus pattern, S_2, the organism will, as we say, '*come to expect*' a stimulus belonging to the stimulus class S_2 whenever a stimulus of the stimulus class S_1 is encountered. This way of talking has been traditionally dismissed by radical behaviourists as unacceptably mentalist - useful no doubt for the purposes of popular exposition, but not to be taken seriously in a scientific account of behaviour. I believe that this view is mistaken. What is wrong with mentalistic language, or rather with *some* mentalistic language, is that it presupposes an organism whose behaviour is controlled by a linguistic formula or "rule" as Skinner calls such things, which "specifies" the relevant contingencies. The trouble with construing behaviour in this way is that 'rule-initiated behaviour', as I suggested we call it in my paper this morning, is

(a) confined to the behaviour of human beings after they have acquired a considerable measure of linguistic competence.

(b) restricted to the initiation of a relatively small proportion of the behaviour of even the most rational of human adults.

(c) a form of behaviour which is built up from and superimposed on behaviour that is, to use Skinner's term, "contingency-shaped".

To speak of an organism learning to 'expect' or 'anticipate' stimulus event S_2 when presented with stimulus event S_1 does not bring with it any such implication that the behaviour in question is 'rule initiated' in this sense. Moreover given a network such as that illustrated on Figure 3, it is not difficult to see what this circuit's 'expectation' or 'anticipation' of S_2 might be supposed to amount to in terms of the microstructure of the brain.

In Figure 3 the expectation or anticipation of S_2 aroused by S_1 is represented by the excitation arriving at the three nodes in the hidden layer nodes otherwise fired by S_2 from nodes in the output layer triggered by S_1 *via* the recurrent or re-entrant circuits.

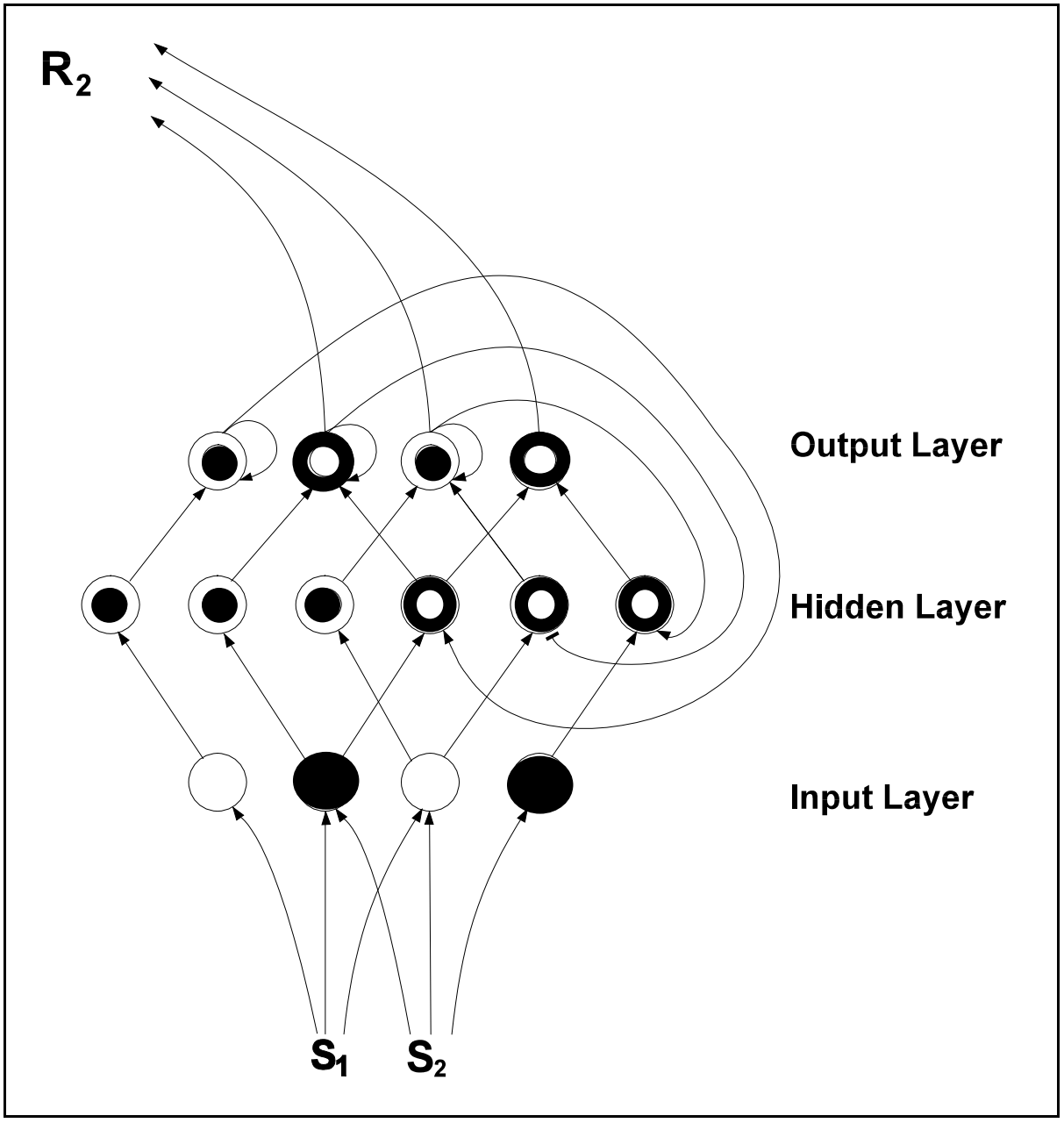

FIGURE 3. NETWORK REQUIRED FOR THE ASSOCIATION OF SUCCESSIVE STIMULUS PATTERNS

This feedback excitation from the output layer to the hidden layer is kept going in a case where there is an interval between the offset of S_1 and the onset of S_2 by activity in the reverberatory circuits shown as feeding out of and immediately back into the relevant nodes in the output layer.

It is assumed that the effect of the simultaneous arrival at the S_2 controlled hidden layer nodes of excitation both from S_2 and from S_1 *via* the recurrent/re-entrant circuits is to strengthen the weights of the synapses connecting S_1 to the response previously triggered by S_2 (Pavlov's "unconditioned response"). When as a consequence of this circuitry, the response is triggered by S_1, it becomes what Pavlov calls "a conditioned response."

As I see it, there are two reasons why the kind of stimulus-stimulus expectancy learning for which Figure 3 provides a model is important for our understanding of the process of concept formation. Not only does it, as Rescorla & Wagner (1972) have proposed, provide what is increasingly recognised as the right account of classical Pavlovian conditioning, the evidence of the reinforcer devaluation experiment (Adams and Dickinson 1981) has led Rescorla (1991) to suggest that instrumental/operant learning needs to be understood in terms of the acquisition by the organism, not of a propensity to respond in a particular way, but of what we may call 'a three-term-contingency expectation.' In other words, what the organism learns is to expect a particular outcome event or *consequence*, given the combination of a particular antecedent stimulus and the sensory feedback from the incipient evocation of a response. This response can then be completed or inhibited depending on the organism's current affective ('pro' or 'con') attitude to the anticipated outcome.

If this analysis is correct, as I am convinced it is, it follows that the process of generalisation and differentiation described by Pavlov in relation to the phenomenon of classical/respondent condition and the formation of stimulus and response classes by the processes of generalization and discrimination learning described by Skinner (1938) are essentially the same process, a learning process for which Figure 3 provides a model.

A second way in which stimulus-stimulus expectancy learning impinges on the phenomenon of concept formation is in providing a background of expectation against which any *unexpected* stimulus sequence will stand out. Whenever such an unexpected sequence is detected, alarm bells ring signalling a problem situation which can only be resolved when the organism's conceptual scheme has been stretched, in the manner which I described this morning, so as to accommodate the unexpected sequence.

So far as I am aware, no connectionist model of attention exists which would allow us to explain what goes on when attention is 'caught', as we say, by the unexpected. What we *do* have, in the shape of Figure 3, is a connectionist model of the process whereby the brain learns what it can safely ignore.

Error-correction learning and the "natural lines of fracture"

About the third and final feature of connectionist networks which help to ensure that the stimulus classes that are formed follow "the natural lines of fracture", I have little to say that I did not say this morning in talking about "Stimulus classes as 'family resemblance' concepts". You may remember that I began that section of the paper by pointing out that

> the common property that needs to unite the members of a stimulus class is the property of reliably signalling the presence of the same *contingency* or set of contingencies.

In other words, what needs to be signalled is the same set of relations between antecedent condition, behavioural response and outcome/consequence. I went on to point out that in order to satisfy this requirement, the organism often needs to form what are, at the level of the stimulus, highly disjunctive stimulus classes with no common pattern or feature running through them. I illustrated this point with the well known example (Figure 4) of the external genitalia of male and female day old chicks which the chicken sexer must learn to use in making the all-important distinction between the two. All that now needs to be added is the comparison between the learning task confronting the trainee chicken sexer and that confronting Gorman and Sejnowski's connectionist network learning to discriminate between mines and rocks by their sonar echo, as illustrated by Paul Churchland in the 1988 edition of his *Matter and Consciousness* from which Figure 5 is taken. It is clear that the features of the sonar echo which distinguish mines from rocks are just as obscure, just as disjunctive, as are those which distinguish the external genitalia of male day old chicks from those of their female counterparts. In both cases, the only way the network in the one case and the trainee chicksexer in the other case can learn to make these vital discriminations, is by the process of random trial and error correction or 'supervised learning', as it is called in the case of the network, 'instrumental/operant discrimination learning' or 'contingency-shaping' as it is called in the animal learning literature.

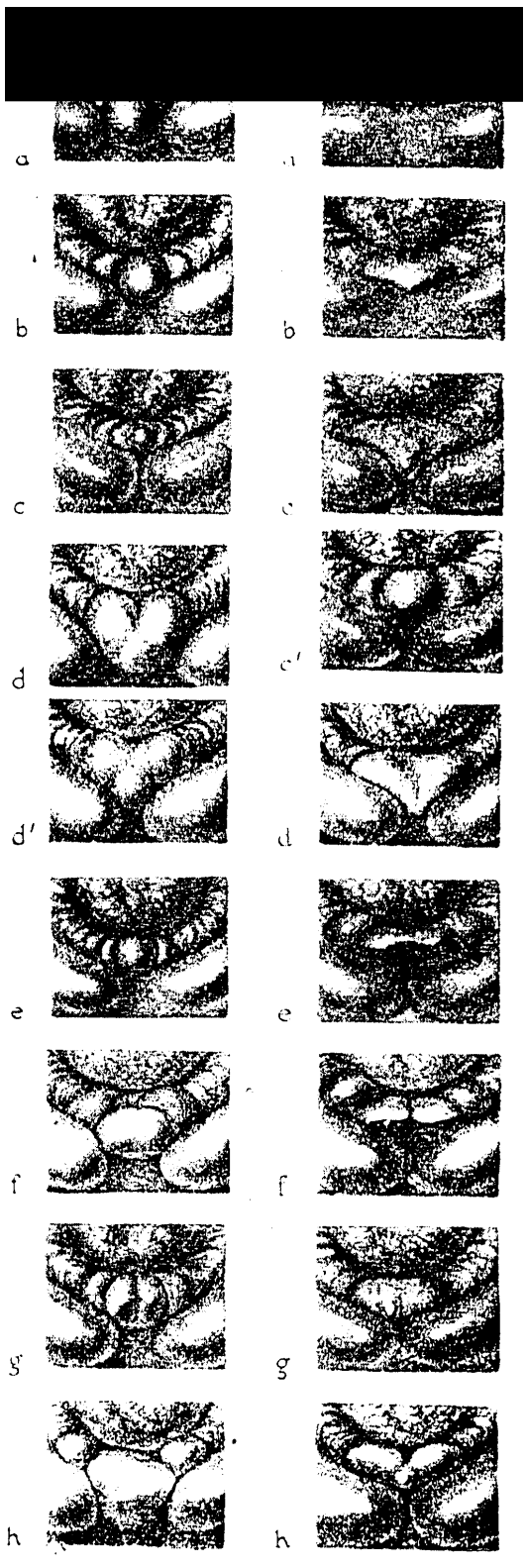

FIGURE 4. External genitalia of male (left) and female (right) day-old chicks (*Canfield, 1941*).

As I indicated in the paper I presented at last years conference, I am inclined to think that the process of 'back-propagation' whereby error-correction is implemented in most connectionist networks, is not the way it works in the living nervous system. This however, does not affect the principle whereby discriminations which follow 'the natural lines of fracture' can only be achieved by exposing the network to a feedback whose form is tightly constrained by the actual and motivationally significant consequences of behaving in one way rather than another.

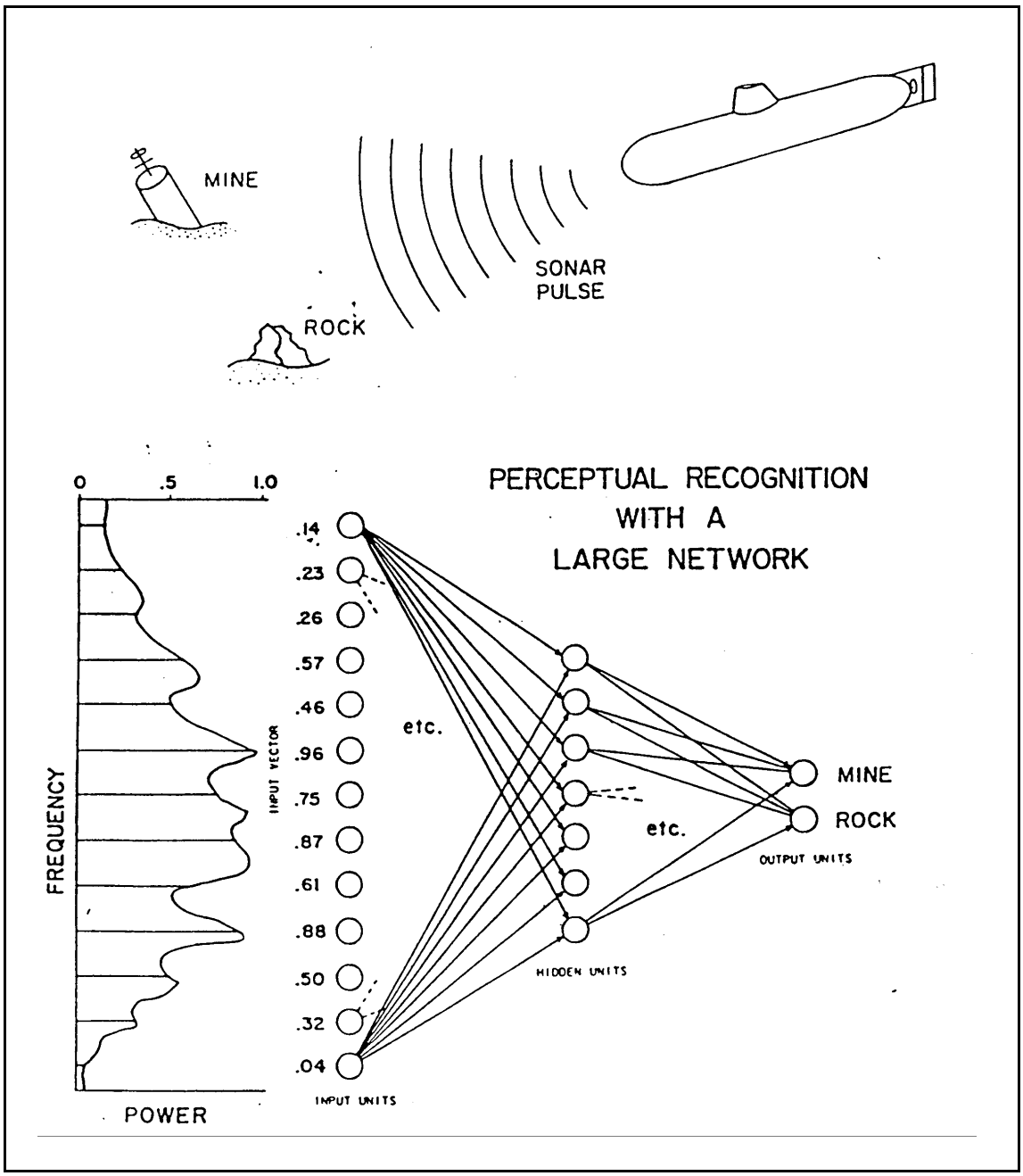

Figure 5. (*From Churchland, 1988, after Gorman and Sejnowski*).

References

Adams, C. D. and Dickinson, A. (1981) Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, **33 B**: 109-112.

Canfield, T. H. (1941) Sex determination of day-old chicks, II. Type variations. *Poultry Science*, **20**, 327-328.

Churchland, P. M. (1988) *Matter and Consciousness*, Revised Edition, Cambridge, Massachusetts: MIT Press.

Edelman, G. M. (1987) *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Fechner, G. T. (1860) *Elemente der Psychophysik*. Leipzig: Breitkopf & Härtel. English translation of Volume I as *Elements of Psychophysics, Volume I* by H. E. Adler. New York: Holt, Rinehart and Winston, 1966.

Jordan, M. I. (1986) Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Meeting of the Cognitive Science Society*. Hillsdale, N.J.: Erlbaum.

Pavlov, I. P. (1927) *Conditioned Reflexes*. English Translation by G. V. Anrep. London: Oxford University Press.

Rescorla, R. A. (1991) Associative relations in instrumental learning: The eighteenth Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology*, **43B**: 1-23.

Rescorla, R.A. and Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black and W. F. Prokasy (eds.) *Classical Conditioning*, Vol. 2: *Current Research and Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Skinner, B. F. (1938) *Behaviour of Organisms*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1975) The shaping of phylogenic behaviour. *Journal of the Experimental Analysis of Behaviour*, **7**, 117-120.

Skinner, B. F. (1981) Selection by consequences. *Science*, **213**, 501-504.