Proceedings of Symposia in Pure Mathematics
Volume 60, 1997

Issues in the Mathematical Modeling of Cortical Functioning and Thought

David Mumford

1. Introduction

In the last few decades, there has been an explosion of research on models of the cortex and how intelligent behavior might be achieved on the basis of these models. But to even start modeling in this way, one must make fundamental assumptions about the biological realities, about thought and about how the two are linked. It seems to me that many of these assumptions are neither self-evident nor definitively established. Therefore, committing yourself to a theory in this type of research, you run the clear risk that if your assumptions are wrong, subsequent understanding of the problem may make your theories not merely wrong but *irrelevant.* The purpose of this article is to make explicit some of these underlying assumptions and to discuss some of the different choices modelers have made. Unfortunately much research in this area, which goes under the name of 'neural nets' (and related areas of artificial intelligence, computer vision, robotics, etc.), has tended to fragment into distinct schools and this has not afforded the opportunity to step back and say – what really is the range of choices we may have?

To set up a mood of skepticism about our current ideas, I want to start by briefly reviewing some of the history of theorizing about these problems. Theories from previous centuries may seem bizarre and utterly misguided, so one must make an effort to realize that these old theories were nonetheless the product of serious study and deep efforts to integrate each generation's knowledge of biology, psychology and philosophy.

It has, of course, always been clear that the senses were concentrated in the head, but in Greek science, for example, whether the brain or the heart was the most significant factor in thinking was disputed. Thus Hippocrates said

> *"From the brain and from the brain only arise our pleasures, joys, laughter and jests as well as our sorrows, pains, griefs and tears. Through it, in particular, we think, see, hear and distinguish the ugly from the beautiful, the bad from the good, ... the brain is the interpreter of consciousness."*

1991 *Mathematics Subject Classification.* Primary 92C20, 68T01, 92J10, 82B20, 82D99.

The first author was supported in part by NSF Grant DMS-91-21266 and by the ARO Center for Intelligent Control GC-R-319561.

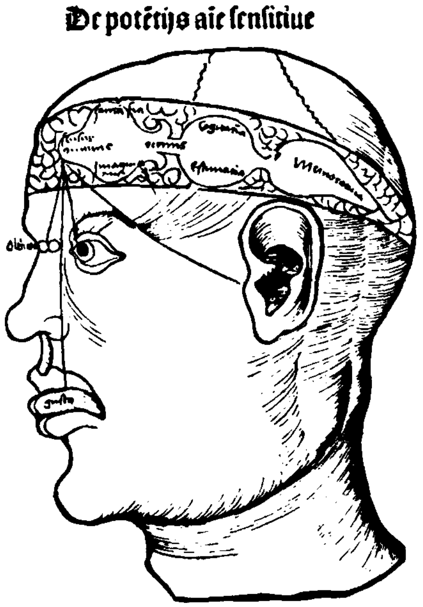

FIGURE 1. Diagram of the three "cerebral ventricles" from Reisch *Margarita Philosophica*, 1513.

But others, including Aristotle, felt that the blood was the essential carrier of thought and that different sense modalities were unified in the heart where you find *"the unitary I who perceives the trumpet"*. Galen, in the 2nd century AD, codified a view which was to be dominant for over a millennium: that the cerebral ventricles were the key structures in the brain correlated to thought and that the nerves were ducts conveying fluids secreted by the brain to the periphery (see figure 1).

In modern times, mind-body duality became one of the dominant paradigms in philosophy and this led to the search for a key organ linking the brain and mind. Descartes believed the pineal gland served this function.

Much more recently, further theories have been proposed which we now believe to contain the germs of truth but which seem in other ways bizarre. One example is Gall's localization theory, in which specific areas of the cortex were believed responsible for many things including generosity, love of parents and secretiveness for example (not to mention resulting in bumps on the cranium when over-developed). Another example was the theory of the Gestalt school of psychology (Wertheimer, Koffka, Köhler etc.) that perception of global form was mediated by a colloidal bio-electric medium supporting 'fields' which interacted thus integrating local and global visual information.

The modern approach to modeling the brain only began with the discovery of the neuron and of the processes linking one neuron to another. An amazing structure was revealed, especially through the work of the brilliant microscopist Ramon y Cajal whose meticulous and detailed renderings of the diverse types of neurons and their dendritic arborizations made it immediately clear that here was a structure whose breathtaking complexity might support something as subtle as thought (see figure 2). Cajal himself was not blind to these implications of his discoveries and said

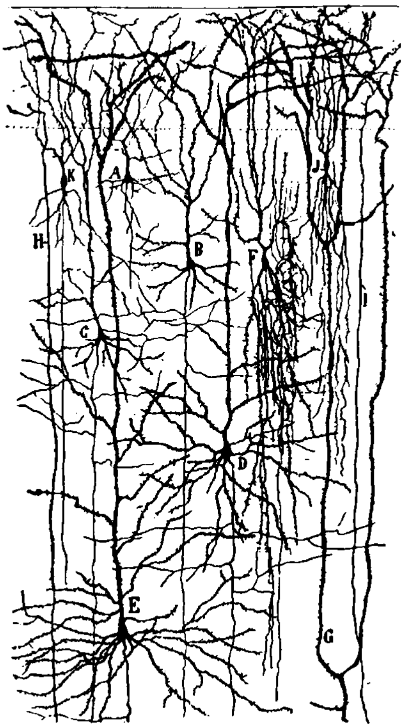

FIGURE 2. Drawing of a Golgi stain of the cerebral cortex of a one month old human infant, from Cajal, 1900.

> *"Like the entomologist in pursuit of brightly colored butterflies, my attention hunted, in the garden of the gray matter, cells with delicate and elegant forms, the mysterious butterflies of the soul, the beatings of whose wings may someday – who knows – clarify the secret of mental life".*

Mathematically, Cajal's work led immediately to the basic proposition that at least on a coarse level, the brain can be modeled by a vast *directed graph* whose vertices were the neurons and whose edges were the synapses from one neuron onto another. Graphs are highly non-trivial mathematical structures and so, on the basis of this observation alone, much theorizing is possible. In the 40's, this led to the first wave of enthusiasm that significant modeling might be possible. This wave was started by the brilliant work of Warren McCulloch **[27]** **[28]**, in which Norbert Wiener participated with the complex of ideas he named 'cybernetics'. But roughly at the same time, however, Sherrington in a dour vein stated his skeptical view that

> *"The relation of mind to brain [is] not merely unsolved by still devoid of a basis for its very beginning."*

Who are we to believe? Will our present day theories of neural nets have someday the same quaint ring as the Gestalt school's or Galton's or even Galen's theories? If this is not to happen, I think we should keep clearly in mind the assumptions that are made by each school of neurobiological and computational theory. In this article, I will focus on three questions which I feel have not been definitively answered, and on which people have taken strong and conflicting stands. The purpose of this article is to present the range of ideas on the basis of which different schools have created theories of the brain and mind. This is an expository article and I will try to avoid taking a stand on which approach seems to me to be correct. Hopefully by focussing on our choices, we can get a better perspective on how far the field has progressed and where to go next, in spite of the hyperbole in the popular press and in the government's grand challenge program announcements.

The three questions I want to discuss are:

1. What is the role of neuronal spikes in carrying information in the brain?
2. What are the "objects" of thought and how are they related to the activity of individual neurons?
3. What does thinking "do" and how do we break it into steps?

2. What is the role of neuronal spikes in carrying information in the brain?

The message conveyed from one neuron to another results from the generation in the first neuron of discrete action potentials or spikes which cause neurotransmitters to cross the synapses and alter the electrochemistry of the second neuron. The pattern of these spikes thus appears to carry information or data in some way from one neuron to another. McCulloch [**27**] [**28**] formulated both the idea i) that a neuron might act as a Boolean device as a result of the arrival of individual spikes synchronously or asynchronously *or* that ii) some stochastic principles might be operating so that individual spikes might not be highly significant. In [**28**], p. 98 he puts it like this:

> *We can summarize our conclusions as follows:*
>
> 1. *the actions of neurons and their mutual relations can be described by the calculus of propositions subscripted for time,*
> 2. *the nervous system as a whole is ordered and operated on statistical principles. Thereby it adjusts the all-or-none laws governing its elements to a physical world of continuous variation*

One can separate recent theories on this question into three camps:

A. The mean firing rate of a neuron carries information; their precise pattern doesn't.

B. Some other aspects of a single neuron's spike train carry more detailed information.

C. Synchrony of spikes or precise lags between spikes of multiple neurons carry information.

(A) is the accepted basis of vast majority of neural net models, as well as the working hypothesis of the majority of neurophysiologists. This is true for a very simple reason: no one knows how to make a neuron produce the same spike train twice in a row[1]. Figure 3 shows 8 neurons each responding to the same experimental situation many times: time runs from left to right and the spike trains of the different runs are graphed as the different closely spaced lines of dots.

[1] However, recently two groups have reported that when driven by complex or more natural stimuli, neurons tend to fire much more predictably than when driven by simplified experimental stimuli [**5**][**15**].

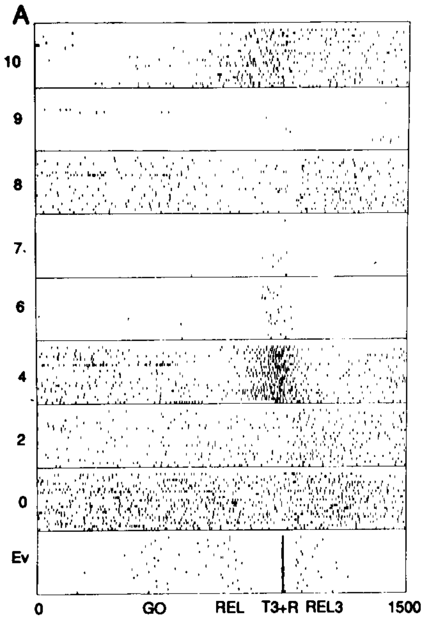

FIGURE 3. The spike trains of 8 neurons to repetitions of the same experiment, showing highly variable responses. From [1].

Note that while some neurons predictably increased their overall firing rate at the corresponding times in the different repetitions of the experiment, there appear to be little or no repetition of individual spikes. For this reason firing rates are considered as the only information transmitted by a neuron. But firing rates may well take 100 milliseconds to measure and this leaves the brain with no time for any complicated calculations with these numbers. In fact, experimentalists, in order to get good comparative data on one cell's firing rate in response to varying stimuli must average the cells responses over many repetitions of the same stimulus. In real life, the brain has no opportunity to have the situation repeated. So along with (A) one must also accept that every signal is actually being conveyed by a population of neurons (say 100) all signalling essentially the same message: then by averaging their responses during a single presentation of a situation, the strength of their signal can be evaluated accurately in e.g. 10 milliseconds by the law of large numbers. Even so, people's ability to respond accurately to complex clues in less than 200-600 milliseconds (these as average reaction times in most psychological tests not requiring reflection) poses a very difficult challenge to a system that takes 10 milliseconds to send each message.

To employ only the mean firing rates of otherwise stochastic spike trains, with the individual spikes being samples from Poisson processes whose mean is set by the sending cell, seems a rather wasteful way for nature to use the exquisite mechanism of spiking and neurotransmitters. People have looked for other possible ways in which information may be encoded in the spike trains. I want to mention briefly some of these ideas which fall under hypothesis B:

1. Richmond and Optican [29] have studied the time course of firing rates of V1 neurons responding to black and white checkerboard patterns (called Walsh

patterns) for some 200 milliseconds after stimulus. They find predictable patterns (such as initially fast, then tapering off; or initially slow, then building up) which depend on the stimulus. But one can hardly call this a 'signal' when it takes 1/5 of a second to transmit: rather they seem to be taking a snapshot of a whole image analysis calculation which is going on during the period of a normal eye fixation.

2. Gray and Singer [**19**] discovered that in anesthetized cats, V1 neurons have a strong tendency to oscillate in what EEG specialists call the γ range: in their case 40-50 Hz. Such oscillations are very widespread in field potentials, but this was the first report of their occurring in single cells. The suggestion was that various populations of cells would synchronize, others not and that in each such population there would now be a clock relative to which individual spikes would have a phase lag, thus carrying more information. The various oscillating populations could be interlaced in time or could entrain each other: the possibilities for theorizing were tremendous. Unfortunately, these oscillations have proven elusive and hard to find in awake primates for example (but see [**26**]).
3. W.Bialek [**8**] has made a close study of the spike trains of two motion detection neurons (responding to motions of opposite sign) in the visual system of a fly and sought to predict, from their spike trains, a complex motion sequence with which the fly was being stimulated. He found that spike timings over an approximately 30 millisecond window following the motion seemed to encode significant information about the perceived motion, though, unfortunately, in a rather untransparent way.
4. Koch and Crick [**23**] have pointed to a specific population of deep pyramidal cells in layer 6 of cortex which burst. Although there is little hard data on the relation of these bursts to stimuli, they theorize that these bursts are highly significant events, possibly correlated to consciousness.

I am sure this is merely a sampling of theories that fall in category (B). This category of hypothesis has the advantage that if something of this sort were correct, it ought to be possible to gather supporting data from single cell recordings. Category (C) raises a much more far reaching challenge to the standard view (A) but is much harder to test. According to this view one must look at the whole *spatio-temporal neural activity pattern* in order to find where information is coded. One should think of this pattern – like the white-noise-like 'snow' on a TV screen tuned to a channel without a signal – as possibly containing spike-level microstructure. A key idea here is that the spikes of an individual cell might be separable into two groups: a stochastic background firing which is bringing all cells up to near firing potentials; and on top of that a small set of volleys of tightly synchronized spikes which are carrying messages and creating equally closely timed responses.

The most developed theory of this sort is the 'synfire chain' theory which Moshe Abeles first proposed about 15 years ago. Like the mean firing rate theory, his theory also requires small populations of say 100 neurons to work together. But instead of simply firing at the same stochastic rate, they give off a single volley of synchronous spikes (synchronized perhaps to within roughly 1 millisecond). What such a volley can do, according to Abeles's calculations, is trigger reliably the firing in further cells to which at least half of the population are connected. If, in fact, the first population is so connected to a second such population, and this to a third,

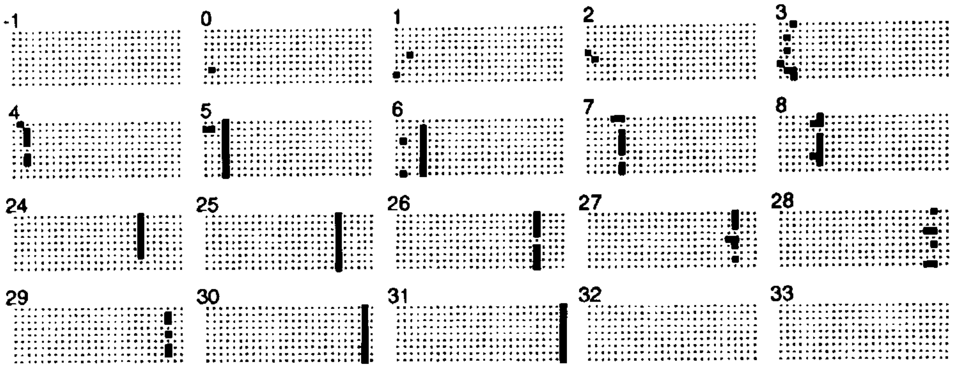

FIGURE 4. Twenty frames from a movie of a simulated synfire chain, where each vertical group represents nine neurons in an assembly, and each assembly is connected to its neighbor on the right. From [**1**].

etc. what he calls a 'synfire chain' is set off. His proposal is these chains are the real carriers of information in cortex. A simulation of such a chain is shown in figure 4 from [**1**].

Indirect evidence for this model comes from Koch and Sofftky [**31**]. They seek to model as accurately as possible the electrochemistry of a single pyramidal cell, including especially its distal synapses, and then to compare the firing patterns that their model would generate with purely Poisson input to actually observed spike trains. They use standard deviation of the interspike interval divided by the mean interspike interval as a statistic to measure spike trains on a scale from very regular to very bursty. The conclusion of their study is that the law of large numbers would make pyramidal cells fire much more regularly if they received purely Poisson input. An alternative possibility which would make the cell output spike trains closer to those actually observed is that these pyramidal cells receive volleys of tightly synchronized input.

After a decade of theorizing, Abeles, partly with Gerstein [**1**] [**2**], has found quite striking data supporting the idea that temporally precise, multi-cellular spiking patterns exist in cortex. Suppose, he says, synfire chains involving some 100 sets of 100 neurons exist in a column with, say, 100,000 cells. Suppose too that hundreds of such chains exist (individual cells belonging to many chains, so the information in this column is multiplexed). Suppose you put a dozen electrodes in this column, picking up at random say 10 cells. He estimates you have a good chance at recording 3 or 4 spikes from some of these synfire chains. These are very small samples from the full activity of the synfire chain, *but* these triples or quadruples of spikes should recur with nearly identical timing whenever the synfire chain is activated. Using a statistical test of what recurring patterns were likely or unlikely to occur by chance, he has found in his experimental data many such excessively repeating coincidences, which he believes to be the traces of synfire chains. Even more striking, he finds these coincidences are concentrated at points where the monkey is making a significant choice or action in his experiment and that, for some cells, most of their activity at these times belongs to such patterns. This is shown in figure 5. These are recordings from an experiment in which the

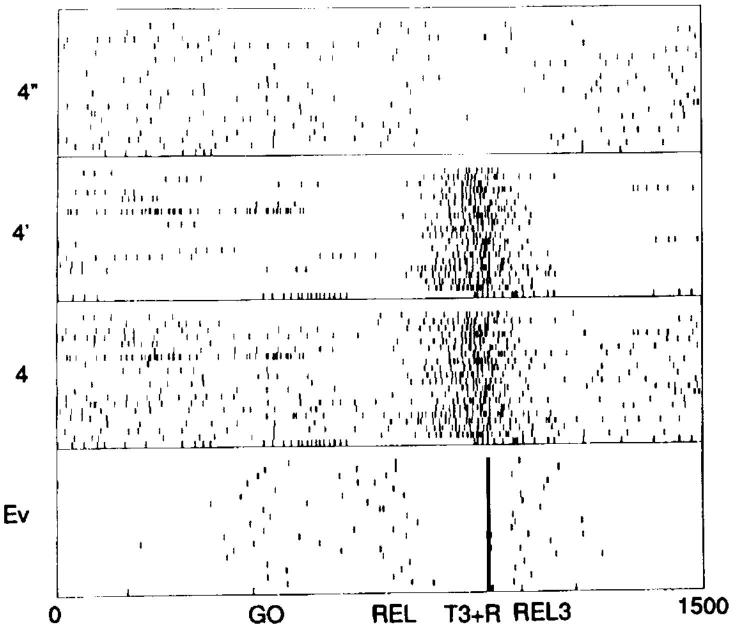

FIGURE 5. Decomposition of the spikes of one neuron (4) into those in excessively repeating patterns (4′) and those not (4″). Note that the patterned spikes cluster around the time marked 'T3+R' when the monkey touched the target and got its reinforcement. From [1].

monkey must observe a clue, then wait and finally, at a prompt, he must touch the right target to obtain juice: the figure shows many recordings from a particular cell, temporally aligned by the instant at which the target is touched: this cell's activity is broken up into the putatively stochastic activity and the spikes which belonged to excessively repeating patterns. Note that the stochastic background is continuous, with some diminution during the touching event, while the patterns give spikes concentrated at the touching event.

Whether or not you accept his interpretation of his data, it is clear that neural net modeling in a brain with synfire chains would have to be totally different from modeling on the basis of spike train frequencies alone. For example, Bienenstock [9] has developed theories of how such chains can organize themselves and how they can be dynamically linked.

3. What are the "objects" of thought and how are they related to the activity of individual neurons?

The second topic I want to discuss is how, given this directed graph of neurons sending and receiving discrete spikes, one imagines computation being carried out by this activity. Specifically, I want to talk about the question of how particular components of thought such as objects being perceived, the identity of these objects, actions being taken and their place in plans, etc. are represented by neural activity. I think it is reasonable to divide the answers to this question again into 3 groups:

A. Grandmother Cell Hypothesis: there is a 1:1 correspondence of concepts and neurons, each neuron with its optimal stimulus.
B. Attractive neural net hypothesis: concepts are stable attractors for the dynamics of local neural populations.
C. Template hypothesis: a concept is a property with slots, instantiated by a temporary assembly of co-firing neurons.

I have stated (A) rather strongly on purpose. In its more usual form, it is phrased by saying that each cell has an *optimal* stimulus or situation in which it will fire most strongly, that it has a graded response to similar stimuli or situations and that for the major components of thought, one should expect to find somewhere in cortex cells which are dedicated to expressing when this is present. This is epitomized by the assertion that somewhere in cortex there should be a cell which fires if and only if you are looking at or thinking about your grandmother – an example which is widely attributed to Barlow, though he points out that it is actually an example due to Jerry Lettvin.

This theory makes cells rather like the fuzzy sets of Zadeh: we can define fuzzy sets of situations, one for each cortical cell, by defining the degree of membership of a situation in each set to be equal to the firing rate of the cell in this situation. There is a dual perspective on this theory: in a specific situation, we expect to find a fuzzy pattern of activity in the neural population, a so-called population coding where one neuron is firing the most and nearby neurons are firing progressively less. In other words, we imagine the correspondence between neurons and objects of thought to be 1:1 but with a fuzzy surround of object–neuron pairs which are more or less similar (a fuzzy mapping between objects of thought and neurons). A major computational activity in this theory is to control the sharpness of the neural representation of a situation: i.e. whether a large distributed pool of neurons is responding in a lukewarm fashion or whether the neural response sharpens to a single neuron firing very strongly. Networks in which the response is sharpened in this way are called 'winner-take-all' circuits. In some situations this is desirable and in others, where uncertainty is present or associations and analogies are desirable, the broader representation may be computationally more effective.

Hypothesis (A) has dominated the neurophysiologist's approach to single cell recordings ever since the ground breaking work of Hubel and Weisel on V1 cells in monkeys and that of Lettvin and Maturana on retinal ganglion cells in frogs. Thus simple and complex cells in V1 typically have a preferred orientation and a graded response to nearby orientations. Georgopoulos [**18**] has found a population coding of the direction of an intended hand movement in area M1 in monkey. A striking example involving more complex percepts is the face and hand cells that Gross and Desimone found in area IT in monkey [**10**]. In figure 6, we reproduce recordings from an IT cell to a series of face and face-like and non-face-like stimuli which illustrates well the idea of a cell having a preferred stimulus and a graded response to similar ones. Other parts of IT have been explored by Tanaka [**32**] using a mechanism of gradually simplifying a complex stimulus to zero in on the essential features which are driving a cell: they find optimal stimuli which are strange combinations of shapes, texture and colors. It does not seem unreasonable to me that the grandmother cell itself might be found in areas such as TF (or TH or TG: medial temporal areas near entorhinal cortex, with connections to cingulate cortex) in the macaque monkey.

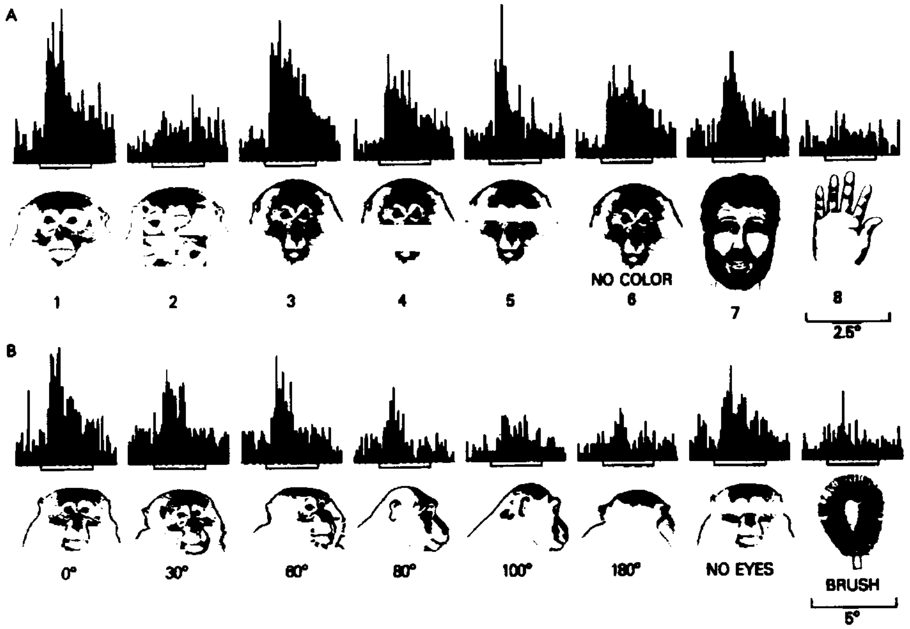

FIGURE 6. Responses of an inferior temporal cell in a Macaque monkey to a variety of faces, jumbled faces, heads and other stimuli. Note the gradual shift from strong responses to optimal stimuli (1 and 3) to no response to a hand or the back of the head. From [**10**].

It seems to me that the most important consequence of (A) is that it makes the objects of thought – let's call them concepts – into a graph too. Is this a reasonable thing? In fact, the earliest attempt to actually do this, to write down *the universal graph of concepts*, is Roget's Thesaurus. Although not widely appreciated by philosophers, I think his creation of the Thesaurus was an inspiration, a truly deep step in describing thinking. Ron Hardin at Bell Labs put his thesaurus on-line and did some striking experiments with it. One such was to investigate the shortest paths in his graph from a concept to its antonym: an example is 'generous' – 'lofty' – 'superior' – 'exclusive' – 'selfish' – 'ungenerous'.

Another linguistic theory which led to a graph of words was the work of Dixon on the Australian aboriginal language Djirbal. This language has 5 classes, like the 2 or 3 genders in Indo-European languages. On first sight, these classes seem to be bizarre, each one containing many pairs of related concepts, but many totally unrelated pairs while other very similar pairs of concepts are separated and placed in distinct classes. Yet the speakers of the language seemed to find the system logical! A striking example was made famous by Lakoff's use of it as the title of his book [**24**]: women, fire and many dangerous things are all part of the class *balan*. Dixon's theory was that this categorization of words in Djirbal was an instance of 'nearest neighbor clustering'. In other words, the aboriginal Djirbal-speaking aborigines had a mental metric of similarity between all the concepts in their world and, as their language took shape, the most similar concepts were seen as having something in common, then related clusters were grouped until the whole universe of concepts was divided up into 5 classes (actually, one class is not like this – it is 'the rest', what didn't fit in anywhere). Statistical runs of the nearest neighbor clustering algorithm on various data sets give very similar results: odd categories with long strung together arms, sometimes coming very close to each other when nearby data points got drawn in opposite directions. My point is that this interpretation of linguistic classes also presupposes a notion of distance between concepts and the idea of linking neighbors in a graph.

Making concepts the vertices of a graph leads to a very beautiful mathematical model of thought and specifically of reasoning about the interpretation of a situation in which – as is always the case in the real world – not every fact is clearly true or false. Start with hypothesis (A) and suppose that the activity of a neuron in a situation corresponds to the *probability* that that concept applies in this situation that the mind is trying to understand. We imagine that the information present is not complete so the mind must make inferences based on past experience about aspects of the situation which are not known. What is needed in order to make such inferences in a mathematically sound way is to have a probability distribution on the space of all possible completely specified situations (where all facts are known to be true or false). Given such a probability distribution, well known statistical principles, such as Bayes's law, enable you to make optimal guesses about the unknown aspects of the present situation. Defining probability distributions on such huge spaces is, in general, totally impractical. But there is a class of probability models known as *Markov random fields* which are practical to define and at least sometimes also practical to reason with. This class requires to start with that its random variables – the concepts in our case – form the vertices of a graph. The definition goes back to Gibbs: given a graph G to whose vertices v we associate random variables X_v, we suppose that we have, for each clique C in G, an *energy* function $E_C(\{X_v\}_{v\in C})$ defined on the random variables in this clique. Then the probability distribution of the set of all random variables is given by:

$$\mathrm{Prob}(X_v = x_v) = \frac{e^{-\Sigma_C E_C(\{x_v\}_{v\in C})}}{Z}$$

where Z is a normalizing constant making the probabilities sum to 1. Modeling particular mental problems by Markov random fields has been a very successful approach, so we state:

- **Extended statistical form of hypothesis A**: The probability model of the world that our brains learn is a Markov random field based on the graph of concepts, which is 'isomorphic' to the graph of neurons.

To illustrate this approach to modeling reasoning and thinking, we reproduce figure 7 from a paper by Lauritzen and Spiegelhalter[**25**]. The figure shows a toy medical expert system which they used to illustrate the principles of a much more complex system they have designed and implemented. The figure shows random variables which correspond to a) possible facts about the prior life of the patient, b) medical facts about the patient's condition and c) symptoms and results of tests. The arrows indicate which variables directly influence which others. This is a slight modification of a Markov random field using a directed graph but the same basic comments apply: by specifying a small set of numbers (the conditional distribution of each variable given the values of its 'parents', the variables connected to it by an incoming arrow) the whole probability space is specified. This approach has also been applied very successfully to speech recognition and to low level vision problems.

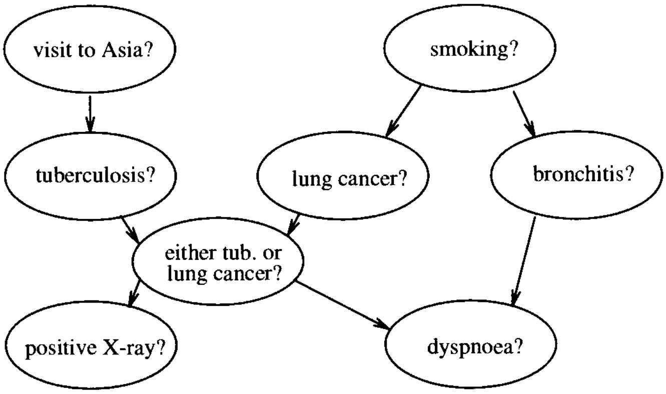

FIGURE 7. The graph of random variables in a demonstration medical expert system based on Markov random fields. Based on [**25**].

The problem which plagues (A) however is how to represent and reason with unforeseen conjunctions of events for which concepts, or neurons, have not been dedicated. A famous example is the yellow volkswagen problem of Barlow: suppose one morning you walk several blocks and pass 10 yellow volkswagens. You probably will begin to notice this odd coincidence and wonder whether it has some explanation! But what leads you to notice this since you never thought specifically about yellow volkswagens before? A similar problem is how to represent mentally a scene containing a red square and a green triangle and distinguish it from a second scene containing a green square and red triangle. Both problems suggest the need to *dynamically bind* two arbitrary concepts for which you have already dedicated neurons in order to represent a new situation. This problem was raised in a well known paper of Fodor and Pylyshin [**14**] challenging the applicability of neural net theory to any but the simplest types of thought and reasoning. Hypothesis (C) gives one possible answer to this.

Another radical answer has been proposed by Valiant [**34**]: that is that every new observed conjunction of stimulus features does indeed find a new cell, which is then dedicated to recognize this conjunction whenever it reoccurs. His premise is that the brain is big and life is relatively short, so one can be profligate in storing many conjunctions. To make this plausible, he makes an elementary observation of the structure of the directed graph of neurons in the brain which, to my knowledge, had never been noticed before. That is, that the *diameter* of this neural graph is 'nearly' 2. Here he is interested not in the maximum over all pairs of neurons of the minimum number of synapses needed to connect them, but in the typical number of synapses for most neuron pairs; and it is acknowledged that the distance in this sense between neurons will be increased by a) the fact that neurons in specific layers tend to synapse on neurons in other specific layers and b) that the brain is divided in to areas with restricted projections from one to another. But if you take neurons in appropriate layers and areas and count numbers, it strongly suggests that for most such pairs of neurons A and B, they can communicate i) by both A and B synapsing on a common third neuron C and ii) by A synapsing on a third C which synapses on B. This high degree of connectivity seems to have great computational significance.

I now turn to hypothesis (B). As opposed to the localized fuzzy concept ideas behind (A), (B) assumes that concepts are represented in a distributed fashion, implicitly through the graph of neurons and its synaptic weights, but explicitly they only manifest themselves in the dynamics of the system. I first heard this theory from E. C. Zeeman [**36**], who developed theories of this type in the 60's. It has been heavily developed in the last decade by Hopfield and Amit and others [**3**][**21**]. We consider the phase space of all possible states of brain activity and look at the propagation of spikes and their generation of further spikes as setting up a dynamical system in this phase space. A concept will then be a *stable attractor* in this dynamical system. It will have a basin of attraction and we can imagine the whole phase space as being partitioned into various such basins. This makes concepts basically Boolean and discrete: the dynamical system cannot fall partly into two such basins of attraction, so the model is closer to classical logic than to fuzzy logic or to probability models. Moreover, while individual concepts in thought are related to the basins of attraction of the dynamical system, the operation of linking two concepts should correspond to altering suitable parameters to push the dynamical system across some kind of bifurcation.

This approach gives rise to the hope that the deep results of dynamical systems theory and of statistical mechanics can be used to analyze the brain. A beautiful example is Amit et al's calculation [**4**] of the number of distinct 'memories', i.e. attractive fixed points, a large Hopfield model (in which the dynamical system is a gradient flow) is expected to have. Unfortunately, these results seem to apply chiefly in the biologically unrealistic case where for every synapse from A to B, there is another synapse from B to A with equal strength. Moreover the behavior of general dynamical systems in dimensions greater than 2 is only beginning to be understood, so bringing in concepts like chaos is considered by some more poetry than science. For example, the use of Grassberger-Procaccia time series analysis to identify strange attractors from EEG recordings has been critically analyzed with negative conclusions by P. Rabb [**33**]. What seems to me the biggest obstacle is how to carry out with fully distributed representations of concepts the kind of difficult inference problem that is addressed by Markov random fields. The great advantage of Markov random fields is that the probability distribution is generated automatically from a sparse set of local dependencies. Otherwise put, Markov random fields work well when there are a large number of pairs of concepts which are conditionally *independent*, when you fix the truth value of a small set of other concepts. It is not clear how to achieve this sort of independence in a fully distributed representation.

Finally, what is meant by hypothesis (C)? Hypothesis (C) starts by rejecting the idea that concepts are atomic indivisible entities. Instead, it supposes that every concept has a structure – a set of components which are linked in various ways in the ideal instance, and that in most actual instances, some of these relations are stretched or some may fail or some pieces may be missing. In other words, a concept is a small graph of other concepts and an instance of the concept is a partial graph match of this ideal or prototype graph with other concepts previously identified. The philosopher's favorite example is the question "Is the Pope a bachelor?" Clearly, the Pope meets some of the requirements to be a bachelor – he is an adult unmarried male – but fails other expectations – he is not a potential husband, nor

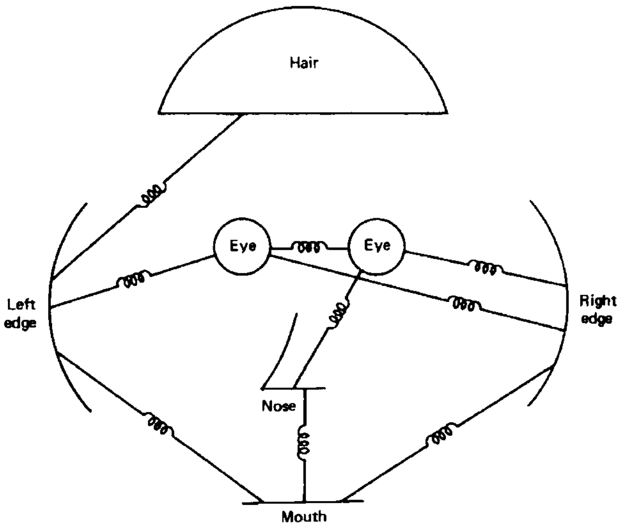

FIGURE 8. To recognize a face, one must identify its parts but allow for individual differences in the geometry of their placement. This was first modeled by Fischler and Elschlager in 1973 [**13**].

does he date. In the neural instantiation of this hypothesis, it is supposed that a set of neurons can be temporarily linked (e.g. on a time scale of 100 milliseconds to several seconds) by synchronized firing or by modification of the effective strength of synapses from one to the other or some other mechanism. NMDA synapses are one possible route for the latter. These ideas have been extensively developed on a cognitive level by linguists such as Lakoff [**24**] and on a neural level by Van der Malsburg [**35**] and Singer [**30**].

For me, the most powerful argument for this hypothesis comes not from higher cognitive thinking but from very simple problems of perception such as the recognition of faces. In figure 8, we reproduce an illustration from a early paper of Fischler and Elschlager on face recognition [**13**]. The idea is simply that a face is made up of several parts with their own individual shapes which must be connected in a certain way, but with individual variations of proportion. To identify a face in the raw visual input, they propose that the location of the various parts must be determined and that one must check that these locations do not distort too much the prototypical shape of faces. I have not seen any plausible alternative to this approach to face recognition, given the huge variation of viewpoint, illumination, expression and facial characteristics that must be allowed for. To put their algorithm in a mathematical form, let $I : D \to \mathbf{R}$ represent the raw visual input (D being the retina or some set of pixels sampling it, I being the intensity of light at each point), and let $I_0 : D_0 \to \mathbf{R}$ represent a prototype face on some standard face-shaped domain D_0. Then identifying a face in the input means finding a diffeomorphism $\Phi : D_0 \hookrightarrow D$ which a) makes the presumed face-like part of the input $I \circ \Phi$ look similar to the prototype face I_0 and b) Φ is a reasonable distortion to expect. (Note that this approach deals with spatial distortions but not variations of lighting.) This can be made into a variational problem by saying that seeking faces in an image is seeking local minima for some functional of the type:

$$E(\Phi) = E_1(D\Phi) + E_2(I_0 - I \circ \Phi).$$

This has a Bayesian interpretation as a maximum *a posteriori* estimate of Φ, if we take $e^{-E_1(D\Phi)}/Z_1$ as the prior probability of a distortion Φ and we take $e^{-E_2(I_0 - i \circ \Phi)}/Z_2$ as the conditional probability of observing an image I, given the presence of a face with coordinates Φ. The key point here is that the random variables that must be estimated are the coordinates Φ of various parts of the face, not the truth or falsity of facts or the strength of some stimulus like redness at a point. One may weaken this demand a bit and, as in the Fischler-Elschlager example, ask only to estimate the coordinates of key points such as the center of the pupils or the corners of the mouth: but 'seeing' a face seems inevitably to entail locating the structures which make up a face. The conclusion is that the concept 'face' does not stand by itself: it has slots – the locations of its key parts – and an instantiation of the concept, at least in the context of perception or mental imagery, requires that these be filled in.

Mathematically, it is hard to fit this into the framework of Markov random fields as discussed in (A). Sometimes this can be done, for instance in time warping for speech recognition, but only because in speech there is a reasonably small number of phonetic events that can be happening at each instant and one can set up a buffer representing each event at each instant of some sample of speech data. In vision, the same approach would require setting up vast numbers of buffers, in registration with the raw input, in which all possible objects to be recognized and all possible parts of these objects were hard wired at each possible position. A different way, which I have been considering, is to set up a class of probability models which I call *mixed Markov models*: these models are distinguished by having two classes of random variables. The first class are the *intensity variables* whose values represent Boolean yes/no truth, or degrees of truth as in fuzzy logic or intensity of activation as in neural nets: these are the usual Markov random field variables. But I want to have a second type of variable, *address variables*, whose values are the labels of other variables. In other words, we have a graph U with two types of vertices, intensity vertices and address vertices, and the value of each address vertex is a second vertex in some subset of the vertices of U. For each assignment of values to the vertices, the address vertices define extra edges in the graph, creating a full graph with some permanent edges and some dynamic edges. The probability distribution is defined by a Gibbs's formula $e^{-\Sigma E_C}/Z$, but with subsets C formed by small sets of neighbors in the augmented graph. This is illustrated in figure 9.

One problem with this hypothesis is what I call the 'shelf of templates' problem. What do you do when you have to recognize 10 faces in a crowd? Or what do you do when a sentence in a novel creates a scene with several families and several parent/child or sibling relationships simultaneously? How can you use a complex template, with slots and prior expectations on relations between the fillers, simultaneously for multiple instances?

4. What does thinking "do" and how do we break it into steps?

The last issue I want to discuss is the subtlest and may sound from the section heading as though it is only a philosophical gloss on more substantive issues. I'm not sure I have found the best way to phrase it, but I feel a major issue is the overall

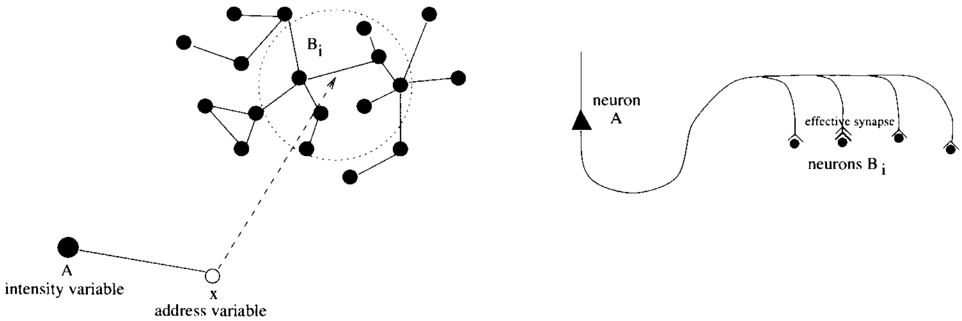

FIGURE 9. On the left, *Mixed Markov models* extend Markov random fields by allowing address variables which add dynamic links to the underlying graph. Address variable x can have any of the nodes B_i as its value. On the right, neural version with 'effective' synapses.

architecture of the system which supports thought and what it is doing second by second. This is so utterly basic to how we model and what we put in our model that it is easily ignored. But, as before, different choices of answers lead you to algorithms which are so unrelated and incomparable that they seem to belong to different fields of research, though they started out with the same goals. Here is my list of four popular answers to this question:

A. Thinking is a process of problem solving, in which you search a tree of possible solutions.
B. Thinking is building a database of facts and rules about the world and deducing their consequences with predicate calculus.
C. Thinking is reacting fast to an evolving unpredictable world with appropriately trained reflexes.
D. Thinking is growing groups of propositions which give consistent, probable scenarios of the ongoing (past, present, future) situation.

Hypothesis (A) I associate with the early days of artificial intelligence, especially with the work of Simon. A paradigmatic example is chess playing, though, of course, the whole point was to apply these ideas to all forms of thinking. Heuristics and what Simon called 'clumping' played major roles. I believe it fair to say that this approach was later recognized as limited and that AI moved towards hypothesis (B), but I include it because it shows one quite distinct answer to the question of what thinking does.

Hypothesis (B) is, perhaps, the mainstream of artificial intelligence, especially in the work of McCarthy. A paradigmatic example for me is the theory of 'naive physics', in which the goal is to codify in predicate calculus our common sense understanding of the world and how it behaves. This turned out to be breathtakingly difficult, much harder than its advocates at first realized. I think the flavor of the approach and a sense of its potential difficulties is very clear from one of the earliest and most influential papers on naive physics, Hayes's paper of the naive theory of liquids [**20**]. We have reproduced in figure 10 one of his figures which summarizes 5 distinct modes of behavior of liquids: you see how tricky it is to codify the many ways in which liquids can behave. I strongly recommend this paper for anyone

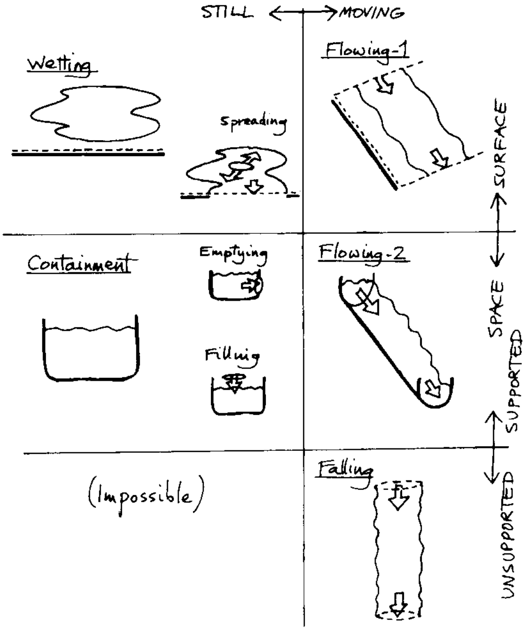

FIGURE 10. Steps in the systematic description by predicate calculus of the many ways in which liquids behave. From [**20**].

wanting to get a bird's eye view of the strengths and weaknesses of mainstream artificial intelligence. Figure 11 shows the characteristic architecture of systems from this school: note that there are peripherals which accomplish 'A-to-D' (analog to digital) transduction of input and 'D-to-A' transduction of output, but within the guts of the system every variable is a clean sanitized Boolean variable. There is always a large database of permanent facts and a smaller one of temporary facts representing the ongoing perceptual and planning situation. One problem with this approach is its dogmatic inflexibility – everything in predicate calculus must be either true or false in spite of the fact that every predicate in the real world seems to have borderline cases. Another is that the things we think about are typically uncertain and we are amazingly good at juggling alternatives and not falling on our faces by accepting blindly one possibility: it is not clear how to incorporate thinking about possibilities into predicate calculus. I will discuss attempts to do this in connection with hypothesis (D) below.

Hypothesis (C) is perhaps closest to Wiener's ideas and Cybernetics: we find a merging of the areas of control theory with the school of artificial intelligence called *reactive planning* (associated especially with Rod Brooks). A paradigmatic example of this view of thinking is the ability to drive a car, and one of the most successful experiments in this direction that of Dickmanns [**12**], who drove a van without human intervention down a German autobahn at 100 km/hr! The emphasis in all this work is on speed: the need to react in real time to an unfolding situation whose unexpected events cannot possibly all be planned for in advance. That is not to say, however, that long term experience doesn't count. This approach places

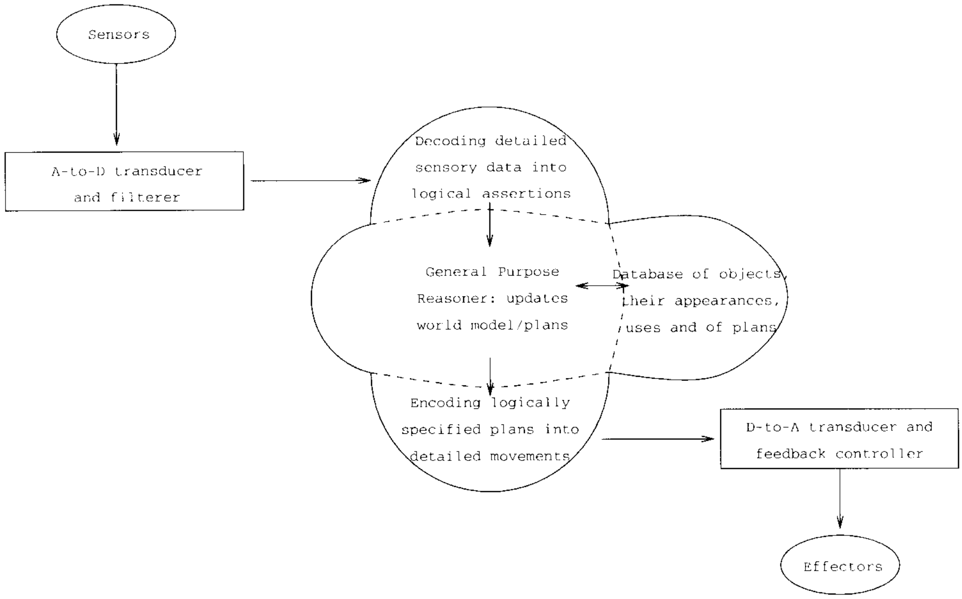
Sensors
A-to-D transducer and filterer
Decoding detailed sensory data into logical assertions
General Purpose Reasoner: updates world model/plans
Database of objects, their appearances, uses and of plans
Encoding logically specified plans into detailed movements
D-to-A transducer and feedback controller
Effectors

FIGURE 11. The characteristic architecture of artifical intelligence systems.

a premium on learning patterns of events from prior experiences, learning what to look for and what may be about to happen next. A fundamental mathematical tool is the Kalman filter which, for linear systems with Gaussian noise, gives an optimal way of merging noisy measurements with an evolving prediction of the state of the world. This filter has been widely applied by linearizing non-linear situations and lumping many other types of unpredictability into the noise term. In figure 12 we show some aspects of Dickmann's work to illustrate this approach: on the top left is his van and next to it an example of the kind of difficult scene in which he must identify the road. The figure below this shows the kind of internal model the system maintains for the road – there is an ideal model for a straight road and a corrected model for what the scene actually looks like if the road curves or the vehicle is not centered and oriented correctly. The last part of the figure shows the overall architecture of his system. Note the arrow labelled 'prediction error' and the box 'Discrepancy interpretation': this is the basic Kalman-filter-like feedback. Also, note the box 'Generation of object-hypotheses' which is the higher level AI modeler. Note that the internal representation of the world is not purely Boolean, but has real-valued 4D spatio-temporal models as well as discrete things like goals.

Another mathematical tool which has been heavily used in this school is non-parametric regression. In the learning phase, it is essential to model accurately how a variable which is not directly observable depends on others which are observed. One seeks to learn algorithms to predict by open-loop (as the control-theorists say) these key variables, because open-loop feed forward algorithms are so much faster than closed-loop algorithms requiring feedback. The algorithms used vary widely: perhaps the most popular is projection pursuit and its neural net variant back-propagation. Others are regression trees and mixture models. Statistical

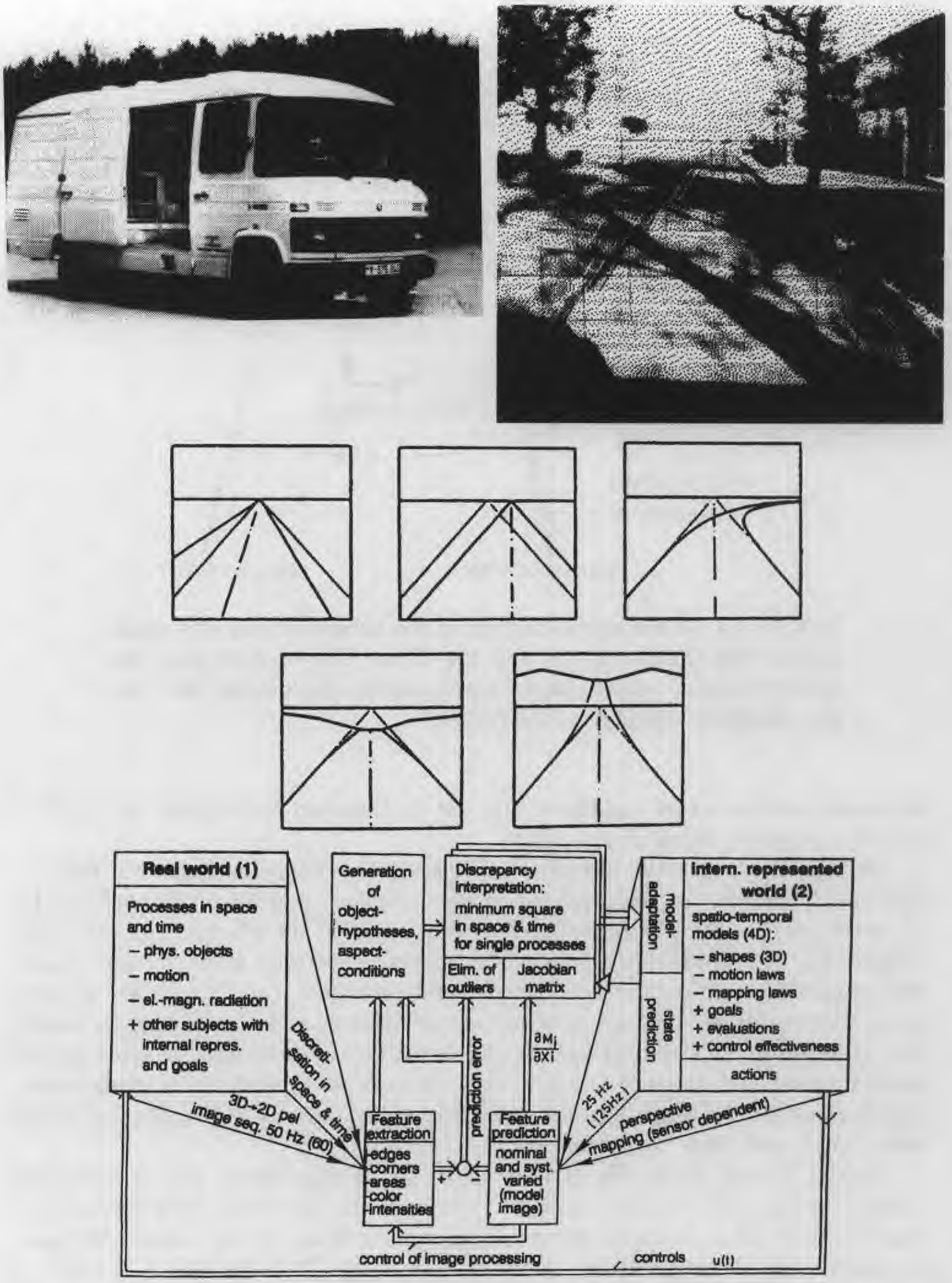
Real world (1)
Processes in space and time
– phys. objects
– motion
– el.-magn. radiation
+ other subjects with internal repres. and goals
Generation of object-hypotheses, aspect-conditions
Discrepancy interpretation: minimum square in space & time for single processes
Elim. of outliers
Jacobian matrix
model-adaptation
state prediction
Intern. represented world (2)
spatio-temporal models (4D):
– shapes (3D)
– motion laws
– mapping laws
+ goals
+ evaluations
+ control effectiveness
actions
Discreti-sation in space & time
3D→2D pel image seq. 50 Hz (60)
prediction error
∂Mi/∂xi
25 Hz (12,5Hz)
perspective mapping (sensor dependent)
Feature extraction
-edges
-corners
-areas
-color
-intensities
+
–
Feature prediction
nominal and syst. varied (model image)
control of image processing
controls u(t)

FIGURE 12. Driving by computer. Top left: The van used by E. Dickmanns. Top right: A scene in which the road is hard to locate. Middle: Fitting internal models to curved and sloping roads. Bottom: The overall Kalman-filter-like architecture of the system. © The MIT Press. E. D. Dickmanns, Expectation based dynamic scene understanding, in *Active Vision*, A. Blake and A. Yuille, editors, MIT Press, 1992.

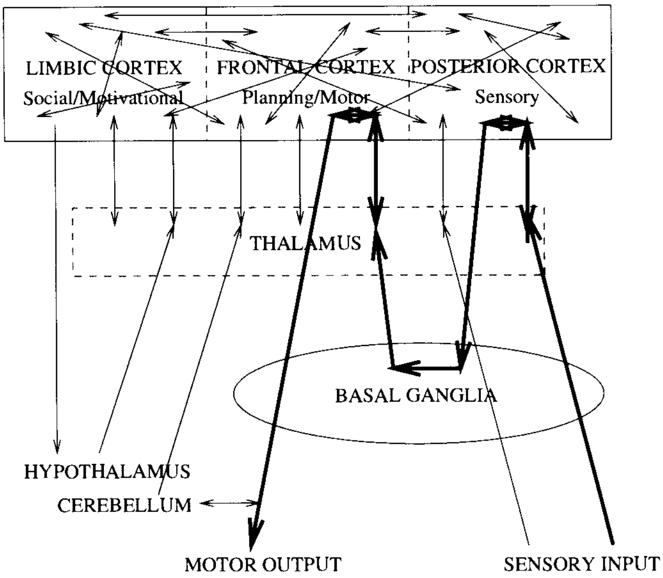

FIGURE 13. A simplified diagram of the neuroanatomy of a mammalian brain, showing in bold the direct flow of data from the senses to motor output via the basal ganglia which avoids the complex feedback circuitry of the cortex.

techniques such as cross validation and use of Bayesian hyperpriors are used in controlling against overfitting the data.

Biologically, it seems as though what this school is modelling is not the full cortical thinking mode, but the much faster 'pre-compiled' mode of acting mediated by the basal ganglia. This biological path is illustrated in the schematic of the brain in figure 13. The idea is that sensory data arrives at posterior areas of cortex where pre-compiled feature extraction is carried out in primary and secondary sensory areas. The results are sent down to finite-state-automata-like basal ganglia, which, after synapses in the striatum and the globus pallidus and thalamic relaying gets to motor or pre-motor areas of frontal cortex. Here pre-compiled motor plans immediately translate the result into motor neuron commands sent down the pyramidal tract: 'quick and dirty' as hackers say.

Finally, I want to discuss (D), which is not as standard a way of modelling thinking as (A), (B) or (C). I consider hypothesis (D) as expressing a key element from the work of a disparate group of people including, among others, MacKay, Grenander and Geman, Minsky, Barwise and Parry. To illustrate it, I want to take as a paradigmatic example the use of the Ising model from statistical physics to accomplish figure/ground segmentation in vision. Recall that the Ising model concerns a two dimensional lattice of iron atoms magnetized either up or down whose energy has two terms: an internal term in which the energy is lowered whenever adjacent atoms are aligned and an external term in which each atom attempts to line up with an external magnetic field:

$$E(\{X_\alpha\}) = \Sigma_{\text{adj. } \alpha,\beta}(X_\alpha - X_\beta)^2 + \Sigma_\alpha (X_\alpha - Y_\alpha)^2$$

where $X_\alpha = \pm 1$ is the spin of the α^{th} atom and $Y_\alpha \in \mathbf{R}$ is the external field. Statistical mechanics is based on the idea that at a temperature T, all states of the iron are possible, but with probability

$$\text{Prob}(\{X_\alpha\}) = e^{-E(\{X_\alpha\})/T}/Z(T).$$

For large T, all states become nearly equiprobable, while as $T \to 0$, only one state remains at all probable, the minimum energy or ground state. The idea of adapting this to combinatorial problems is due to Kirkpatrick [**22**] and the idea of specifically using it for perceptual problems is due to S. and D. Geman [**17**].

Suppose we have an image in which there is a clear separation of foreground and background and the foreground object is more or less dark on a lighter background. Figure 14 is an example, with the dark cow and tree in the foreground and the sky and field to the left in the background. Their idea was to replace the iron atoms in the Ising model by the pixels of the image and to associate one spin, say $X_\alpha = -1$ to pixels in the foreground figure, while associating the other spin $X_\alpha = +1$ to pixels in the background. Then the external magnetic field is replaced by the raw observed image, whose intensity is scaled so that the tones of the foreground are mostly less than 0 while those of the background are mostly greater than 0. The Gibbs model for the physical system now becomes a Bayesian model for any such foreground/background scene. We consider all possible black and white images, representing a black foreground and white background. Most such scenes would look like random black and white dots and would not be thought of reasonable scenes. Instead we suppose that the prior probability of such a scene goes up if there are islands of contiguous black dots and islands of contiguous white space, making a more coherent arrangement. This translates into a prior probability distribution:

$$\text{Prob}(\{X_\alpha\}) = e^{-\Sigma_{\text{adj. }\alpha,\beta}(X_\alpha - X_\beta)^2/T}/Z_1(T)$$

on the set of all possible scenes $\{X_\alpha = \pm 1\}$. Secondly, we suppose that the observed scene is not so simple, but that the intensity of the raw image at pixels in the foreground tend to negative and vice versa in the background. This translates into a conditional probability:

$$\text{Prob}(\{Y_\alpha\}|\{X_\alpha\}) = e^{\Sigma_\alpha (X_\alpha - Y_\alpha)^2/T}/Z_2(T)$$

The final result is that this very simple Bayesian probability model of dark and light foreground/background scenes is identical to the Gibbs probability space of states of the iron crystal.

So far this has rather little to do with hypothesis (D). But now suppose we want to compute the most probable foreground/background segmentation, given the fact that some raw image $\{Y_\alpha\}$ has been observed. What we are looking for is the analog of the ground state of the iron crystal, the minimum energy state. The idea of Kirkpatrick and the Gemans is to use 'simulated annealing': we start the algorithm at a high temperature and we sample hypotheses for the values of the variables $\{X_\alpha\}$ the way nature does with the iron. That is, we make random small changes in one variable at a time based on the probabilities of the states with this variable changed or not. After a while we lower the temperature a bit and let the system move randomly some more but following the new probability distribution in which lower energy has more effect. We continue to lower the temperature until we finally set it to 0. Then the system settles down to a local energy minimum, which, if we went slowly enough, is almost surely the true ground state. Figure

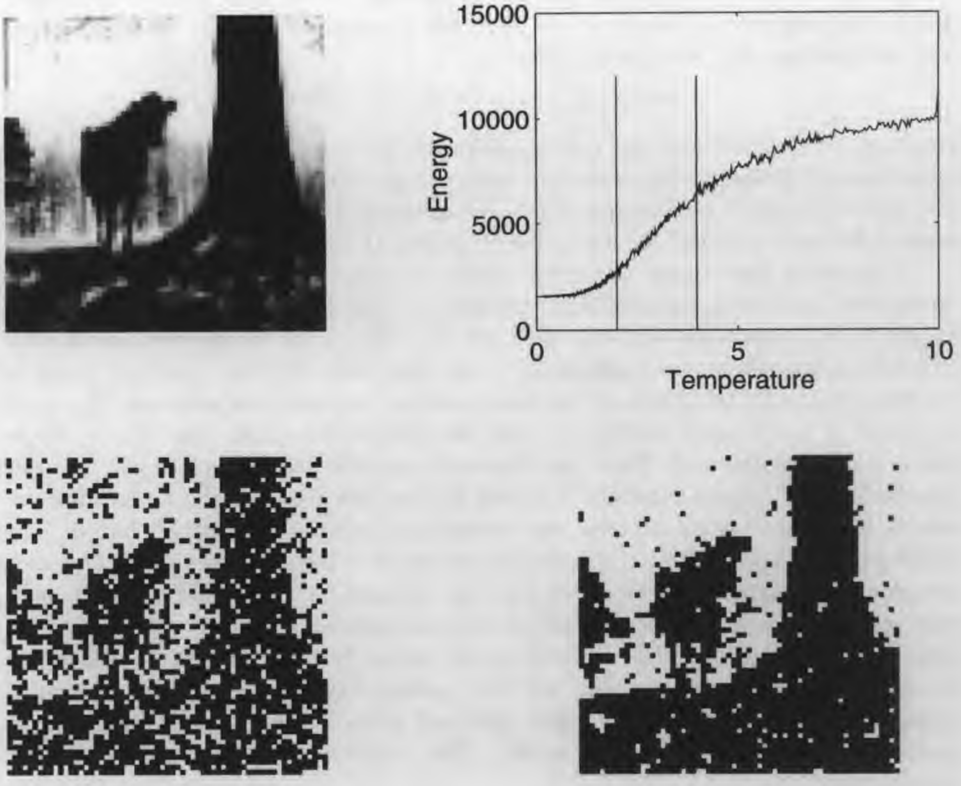

FIGURE 14. Top left: Input image. Top right: Energy as a function of temperature during the annealing process. Bottom left: An intermediate stage at a relatively high temperature in the construction of foreground/background by annealing. Bottom right: A lower temperature stage. The temperatures of two samples are marked by the vertical lines in the graph: note that they delineate the steepest part of the curve – the 'phase transition' in this example.

14 is a result of following this algorithm. It shows an intermediate state, in which the temperature is roughly half way to 0: coherent structures are emerging in the sampled state $\{X_\alpha\}$ which is shown. The graph on the bottom shows how the energy of the overall system decreased as the temperature was lowered. A kind of phase transition is occurring in which the energy is decreasing faster at this temperature because coherent shapes are located. The essence of this algorithm is *to seek local coalitions of consistent data, gradually enforcing stronger and stronger long range ties until a maximally mutually consistent interpretation of the scene is arrived at.*

This extremely concrete elementary example may seem to have little to do with thinking in general. To see the relationship, imagine we have a very large number of possible objects, processes and relations in some situation in the world that we are trying to think through. The medical expert system example discussed above, see figure 7, shows the kind of variables that might be involved, although it deals with a tiny number of variables. Usually we know the truth or the values of some of the variables well, others we have some guesses about. Usually each variable is directly connected to some set of 'neighbors' exactly as in the Markov random field model. In this case, what the Ising model example calls for is to explore possible values for the unknown variables, first on the basis of what values are most compatible with immediate neighbors, then seeking hypotheses on the values of larger groups of variables which are maximally consistent with the data, etc. I hope this makes clear the analogy with the simple Ising model. Simulated annealing is a rather extreme way of seeking the most probable set of values of the unknown variables, called the *maximum a posteriori* estimate of their values. Other algorithms may be more efficient but one must always expect that multiple possibilities must be looked at, that there is no simple way to zero in immediately on the one best guess. In statistical terms, what we must do is *sample the posterior*. In addition to simulated annealing, other techniques such as genetic algorithms, split and merge algorithms, deterministic annealing, auxiliary variable methods, etc. have been proposed [**7**].

It might seem that this type of approach to thinking is about as far as one could get from the logic based ideas in (B). In fact, there has been a steady development of more and more radical modifications of classical logic which have pushed it closer and closer to approach (D). The biggest problem of applying logic to the world is that in the real world there is never one absolute truth. Thus modal logic was developed to express possible truths, temporal logic was developed to express the truths of the past and future, epistemic logic was developed to express the view of the world in other people's thoughts and non-monotonic logic was developed to express tentative conclusions based on assumptions which hold in typical situations. In these logics, there was a parallel development of the syntax – the formal structure of assertions within the theory – and of the semantics – the metatheory within which truth values can be determined. From Kripke's first work on modal logic, the key to their semantics was to consider not just one true and eternal world but a set of possible worlds: the worlds which different agents believe in, or the worlds of the past and future or possible worlds which just don't happen to be the same as ours. For example, Montague [**16**] analyzed the idea of possible worlds with the goal of expressing why the sentence 'John seeks a unicorn' does not mean the same thing as 'John seeks a member of the empty set'.

But the most radical step in this direction is *situation semantics*, a theory invented by Barwise and Perry [**6**], [**11**]. Here possible worlds are elevated from lurking in the background in the semantics to a full-fledged syntactical role and renamed 'situations'. I believe the following is what is meant by a situation:

- A *situation* is a set of compatible assertions,
  - some listed explicitly,
  - some implicitly by reference to actual sensory data or to potential experiences and queries (even ones you can't carry out, e.g. what Caesar and Cleopatra said to each other),

  which can be located
  - anywhere in space-time,
  - in the minds of people (as beliefs, desires or plans),
  - or may be purely hypothetical, mythical or even counter-factual ("If JFK had not been shot, then ...").

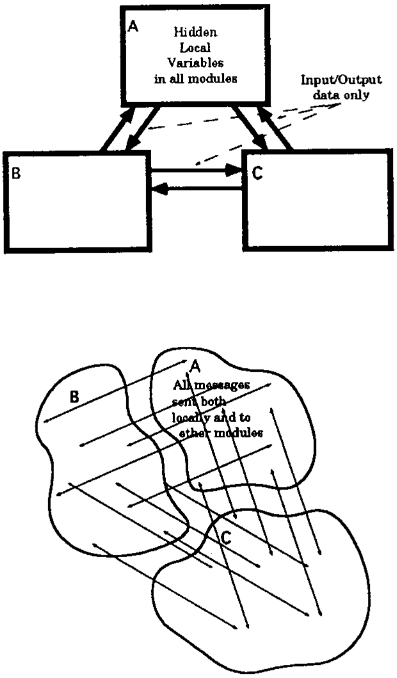

FIGURE 15. Two modes of organizing a computation. Top: A strictly modular approach in which steps are separated in black boxes, exchanging limited well-specified data. Bottom: A public approach in which all data is shared and different modules iteratively seek compatible conclusions. The latter is charactersitic of the mammalian brain.

Except for the absence of probabilities, a situation plays the same role in this logical theory as what you get by sampling the posterior in the statistical approach. In both cases, *we are seeking to construct a hypothesis by growing larger and larger groups of assertions which are compatible and which support each other.* In both cases, many hypotheses are considered in the process of thought which will ultimately be discarded.

Hypothesis (D) has an interesting biological side too. It has been known for a long time that the cortex is subdivided into areas each of which is strongly connected to itself by the local collateral branches of the axons of its pyramidal neurons. *Between* these areas, there is a much more restricted pattern of pathways which travel down through the white matter along the principal branches of the axons of the pyramidal cells. This architecture is consistent with a modular decomposition of brain functioning, quite parallel to the modular programming style which has been developed in order to handle complex computer code. This modular computer style of algorithm is based on subdividing a problem into more or less independent steps and having more or less independent subroutines (with their own private local variables) do each step, only communicating their essential input and output. However, the cortex confounds this neat picture because of several facts: first, whenever one area, call it A, projects its axons to another area B, then it turns out that B also projects back to A. And secondly, about two-thirds of neurons within every area project both locally and to an external area. This is illustrated in figure 15. In other words, the areas have no privacy – every aspect of their activity is being sent to some other area – and every message seems to generate a reply. Rather than thinking of modular computer code, a better analogy might be a relaxation algorithm, in which successive iterations of some set of update rules continue until equilibrium is reached. It suggests that the cortex is 'growing' a larger and larger set of mutually supportive propositions, individually expressed by a pattern of activity in a specific area of cortex where the messages from one area to another seek to strengthen the propositions in the second area compatible with those in the first. This is hypothesis (D).

References

[1] M. Abeles, Dynamics of Neuronal Interactions in the frontal cortex of behaving monkeys, *Concepts in Neuroscience*, **4**, 131-158, 1993.

[2] M. Abeles and G. Gerstein, Detecting spatiotemporal firing patterns among simultaneously recorded single neurons, *J. Neurophysiol.*, **60**, 909-924, 1988.

[3] D. Amit, *Modeling Brain Function*, Cambridge University Press, 1989.

[4] D. Amit, H. Gutfreund & H. Sompolinsky, Spin Glass models of Neural Networks, *PhysRev. A*, **32**, 1007-1018, 1985.

[5] W. Baer & C. Koch, Precision and reliability of neocortical spike trains in the behaving monkey, *The Neurobiology of Computation: Proc. 3rd Comp. and Neural Systems Conf.*, Kluwer, 53-58, 1994.

[6] J. Barwise and J. Perry, *Situations and Attitudes*, MIT Press, 1983.

[7] J. Besag & P. Green, Spatial Staistics and Bayesian Computation, *J. Royal Stat. Soc. B*, **55**, 25-37, 1993.

[8] W. Bialek, F. Rieke, R. Steveninck & D. Warland, Reading a Neural Code, *Science*, **252**, 1854-1857, 1991.

[9] E. Bienenstock, A Model of Neocortex, Brown Univ. Dept. of Appl. Math. technical report, 1994.

[10] R. Desimone, T. Albright, C. Gross & C. Bruce, Stimulus-selective Properties of Inferior Temporal Neurons in the Macaque, *J. Neuroscience*, **4**, 2051-2062, 1984.

[11] K. Devlin, *Logic and Information*, Cambridge University Press, 1991.

[12] E. D. Dickmanns, Expectation-based Dynamic Scene Understanding, in *Active Vision*, A.Blake & A.Yuille editors, MIT Press, 1992.

[13] M. Fischler & R. Elschlager, The Representation and Matching of Pictorial Structures, it IEEE Trans. on Computers, **22**, 67-92, 1973.

[14] J. Fodor & Z. Pylyshin, Connectionism and Cognitive Architecture, *Cognition*, **28**, 3-71, 1988.

[15] J.L. Gallant, C.E. Connor & D.C. Van Essen, Responses of visual cortex neurons in a monkey freely viewing natural scenes, *Soc. Neuroscience Abstracts*, **20**, 1054, 1994.

[16] L. T. F. Gamut, *Logic, Langage and Meaning: Intensional Logic and Logical Grammar*, Univ. Chicago Press, 1991.

[17] S. and D. Geman, Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images, *IEEE Trans. Patt. Anal. Mach. Int.* , **6**, 721-736, 1984.

[18] A. Georgopoulos, J. Lurito, M. Petrides, A. Schwartz & J. Massey, Mental Rotation of the Neuronal Population Vector, *Science*, **243**, 234-236, 1989.

[19] C. Gray & W. Singer, Stimulus-specific Neuronal Oscillations in the Cat Visual Cortex, it Proc. Nat. Acad. Sci.}, **86**, 1698-1702, 1989.

[20] P. Hayes, Naive Physics I: Ontology for Liquids, reprinted in *Formal Theories of the Commonsense World*, ed. J. Hobbs and R. Moore, Ablex Press, 1985.

[21] J. Hopfield, Neural networks and physical systems with emergent computational abilities, *Proc. Nat. Acad. Sci.* , **79**, 2554, 1982.

[22] S. Kirkpatrick, C. Geloti & M. Vecchi, Optimization by Simulated Annealing, *Science*, **220**, 671-680, 1983.

[23] C. Koch & F. Crick, Some Further Ideas Regarding the Neuronal Basis of Awareness, in *Large Scale Neuronal Models of the Brain*, C. Koch & J. Davis editors, MIT Press, 1994.

[24] G. Lakoff, *Women, Fire and Dangerous Things*, Univ. Chicago Press, 1987.

[25] S. Lauritzen and D. Spiegelhalter, Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems, *J. Royal Stat. Soc. B*, **50**, 157-224, 1988.

[26] M. Livingstone, Visually Evoked Oscillations in Monkey Striate Cortex, *Neuroscience Abstracts*, **17:73.3**, 1991.

[27] W. McCulloch & W. Pitts, A Logical calculus of the Ideas Immanent in Nervous Activity, *Bull. of Math. Biophysics*, **7**, 89-93, 1943.

[28] W. McCulloch, The Statistical Organization of Nervous Activity, *Biometrics*, **4**, 1948.

[29] B. Richmond & L. Optican, Temporal Encoding of 2-dimensional Patterns by Single Units in Primate IT Cortex, *J. Neurophysiology*, **57**, 132-178, 1987.

[30] W. Singer, Putative Functions of Temporal Correlations in Neocortical Processing, in *Large-Scale Neuronal Theories of the Brain*, edited by C. Koch and J. Davis, MIT Press, 1994.

[31] W. Sofftky & C. Koch, The Highly Irregular Firing of Cortical Cells in Inconsistent with Temporal Integration of Random EPSP's, *J. Neuroscience*, 1993.

[32] K. Tanaka, Neuronal Mechanisms of Object Recognition, *Science*, **262**, Oct.29, 1993.

[33] J. Theiler & P. Rapp, Re-examination of the evidence for low-dimensional nonlinear structure in the human EEG, preprint, Dept. of Physiology, Medical College of Pennsylvania, 1994.

[34] L. Valiant, *Circuits of the mind*, Oxford Univ. Press, 1994.

[35] C. Von der Malsburg, The Correlation Theory of Brain Function, Max Planck Institut for Biophysical Chemistry, Report 81-2, 1981.

[36] E. C. Zeeman, Topology of the Brain, in *Mathematics and Computer Science in Biology and Medicine*, Medical Resarch Council, London, 277-292, 1965.

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY, CAMBRIDGE, MA 02138
E-mail address: mumford@math.harvard.edu