What Is the Computational Goal of the Neocortex?

Horace Barlow


INTRODUCTION

The human species originated very recently and has been changing very rapidly. Since the neocortex is the main structure that enlarged in primates and now makes us (for our body size) the biggest brained of all animals, its selective advantage is probably responsible for this extraordinary rapid evolution. Figure 1.1 attempts to give a perspective on all this by displaying the history of our species on a cosmic time scale, and it shows both that our status has been changing at a breathtaking rate over the past 10,000 years, and that there is now a serious threat of overpopulation of the earth by humans. Does this mean that the neocortex has done its job too well? And if it has, is there any alternative to further trust in its supposed product—rational action planned by rational thought—to avert the overpopulation threat? How the neocortex evolved so rapidly and what it does are important problems.

This chapter starts by emphasizing the inadequacy of the historical account of the evolution of the human neocortex, and the insufficiency of the neurophysiological account of it as providing a processed representation of the current sensory input. Next a role for it is suggested that combines and reconciles the neurophysiological view with that of comparative anatomists, who have told us that it acquires and stores knowledge of the world. At first these views appear to be quite different, but the hypothesis that the neocortical representation is specialized to facilitate the identification and learning of new associations amalgamates them. The middle part of the chapter sets out the requirements for such a specialized representation, and it is shown that a working model or cognitive map of the world is entailed in its production. This map or model would be used automatically in representing sensory information, but the knowledge that the code embodies might also be accessible by a different route for imagery and recall. I think the hypothesis provides a new and illuminating way of looking at the key role of perception in mediating between sensation and learning. The last part of the chapter outlines collaborative work, still incomplete, prompted by the hypothesis and done with A. R. Gardner-Medwin and D. J. Tol-
Figure 1.1 The top line shows the prominent events from the big bang to the present day on a linear time scale. The next scale shows the prominent events in the last 1/100th of this enlarged 100 times, and so on for the next two scales. The final scale also enlarges the final 1/100th, but places 1990 at the center. The shaded area under the curve gives total human population estimated from Thacher (1983). It has currently reached a density of 25 per square mile, averaged over the whole surface of the earth, oceans, arctic wastes, and all Human culture is very recent and has been accompanied by an explosive growth in human population.
The two questions were (1) How easy is it to identify the association of reward or punishment with the logical conjunction of two or more active representational elements? This is the "Yellow Volkswagen" problem posed by Harris (1980); Gardner-Medwin has shown that this can be done with reasonable efficiency in the case of frequently occurring conjunctions in sparse representations, but rare conjunctions in dense distributed representations will be masked by noise resulting from accidental associations with the separate constituents of the conjunction. (2) What features should be directly represented by single elements in order to promote the efficient identification of associations? It is argued that one should choose as primitives conjunctions of active elements that actually occur often, but would be expected to occur only infrequently by chance. Tolhurst has done measurements on natural images confirming that edges, which the brain certainly does use as representational elements, are aptly described as such "suspicious coincidences."

Inadequacy of the Historical View of Cortical Evolution

A historical explanation is usually given for our large brains and intellectual domination of the world. Our earliest mammalian ancestors, it is said, were ground-living creatures with smell as their dominant sense, but when they colonized the trees smell became less useful, whereas sight, sound, and muscular dexterity became more important. Smell formerly dominated the forebrain, and when it lost its importance this freed the protocortex for other purposes, so the small regions previously devoted to vision, hearing, and touch and muscular movement rapidly expanded and thus formed the primitive neocortex. This organization enabled our ancestors to expand into new ecological niches, and the improved associative power of the new organ gave us the intellectual advantages, including versatility, insight, and adaptability, that have enabled us to dominate the world. The outline of this view dates back at least to Elliot Smith (1924), but many details have been added (Allman 1987, Jerison 1991).

This crude sketch does not do justice to several nice aspects of this story, but it is basically unsatisfactory because the neocortex appears to have led the evolution of mammals, primates, and man, and not to have followed passively as a result of a series of historical accidents. What selective advantage could the forebrain, or future neocortex, provide that other brain regions could not? What is meant by improved associative power, and why should an organ formerly dominated by smell have it? These are the interesting questions, and the history of man's evolution is not the right place to look for the answers.

The supposed origin of neocortex in a region specializing in olfaction is interesting, but that is a difficult fact to interpret and would not make a good starting point. Instead we look at the account of neocortex that neurophysiology has given us.
Inadequacy of the Neurophysiological Account of Neocortex

As many have recognized, the view of cortical function derived from neurophysiology is unsatisfactory. Something like 60% of the monkey cortex seems to be directly connected with vision (Van Essen and Maunsell 1980), but so far no one has really tried to understand how it does anything but represent the current visual scene. The same is true in other modalities—it is the function of representing the current input that has received attention. But we do not have a homunculus to look at these representations: our cortex and associated structures form the representation, look at it, analyze it, store results about it, use it, and continuously add to it. Animals learn almost everything they know through their senses, and academic knowledge apart, the same is true for us; but the means of acquisition, storage, and utilization of this knowledge have been little thought about or studied, and I think it is time to accept that the neocortex must do more for us than merely represent the current scene.

A Hypothesis about the Computational Goal of Neocortex

The outstanding question is: How does neocortex give the great selective advantage that must lie behind our rapid evolution? Herrick (1926) said that the cerebral cortex provided the "filing cabinets of the central executive," and he also called it the "organ of correlation." Jerison (1991) summarizes its role as "knowing about the world." The importance for higher mental function of forming working models and cognitive maps of the world was pointed out by Craik (1943) and Tolman (1948), and although heretical at that time—these ideas from psychology fit the view from comparative anatomy very well and are now widely accepted. Tolman was thinking primarily of representing the geographic layout of the world, and Craik’s working models imitated the dynamics of interactions in the material world, but as Humphrey (1976) pointed out, the interactions between people are the most complex and important things we have to understand, and the cortex is therefore likely to be much concerned with this aspect.

Thus the hypothesis is that the cerebral cortex confers skill in deriving useful knowledge about the material and social world from the uncertain evidence of our senses; it stores this knowledge, and gives access to it when required. This extremely complex and difficult task specifies a definite computational goal for neocortex, providing a useful framework for thinking about its structure, organization, and function. First consider the problem of acquiring such knowledge.

Information and Knowledge

We understand the problem of acquiring knowledge of the world better now than in Herrick’s day. It is not a matter of simply recording or video-
taping the succession of messages from the outside world that our senses provide, but is a much more analytic process. For present purposes it is convenient to distinguish two aspects of the stream of sensory data, information and knowledge. Information is unpredictable, both from previous parts of the stream of data and from other parts of the current stream. As Shannon told us (Shannon and Weaver 1949), all this genuine information can in principle be encoded on to a channel of much lower capacity than that which is required for the physical data transduced by the sense organs. The structure and regularity in the stream of data are redundancy in terms of information theory, but this part constitutes the knowledge that the neocortex must continuously acquire and use. Both parts, information and knowledge, are important to the brain: it must recognize the structure and regularity both to distinguish what is new information and to make useful interpretations and predictions about the world. Finding the structure and regularity is the analytic part of dealing with the succession of sensory impressions that the brain receives, and this is the part that the neocortex performs better than other brain structures according to the current hypothesis: it gives meaning to the stream of sensory data.

The Salience of Structure

Look for a moment at the top left part of figure 1.2. It consists of a random array of dots. Compare it with the top right, where each dot has been paired at a position symmetric about the center line. The random parts of these two figures are identical, but this is not obvious. It is the symmetric structure on the right that leaps to the eye, while the structureless array means nothing to us—unless we look at it long enough and start to impose structure on it, such as faces or other imaginary forms. In the lower two figures the structure resulting from other pairing rules stands out equally clearly, and it is pretty obvious what these pairing rules are. However, it is not at all obvious that a pairing rule is solely responsible for the structure seen: it is hard to believe that the vivid streaks and swirls result just from pairs of data, with no longer concatenations, but that is the case (for the first description of these figures see Glass 1969).

One can detect structures of this sort when they are overlaid by a huge number of completely randomly placed dots (Maloney et al. 1987), so the suggestion is that our perceptual system grabs simple examples of world-knowledge of this sort and uses them to construct its representation of the world. It is plausible to suppose that mirror symmetry and translational symmetry are so abundant in our sensory diet that an animal is certain to encounter them; hence mechanisms for detecting these forms of structure will always prove useful, and their universal provision by ontogenetic mechanisms has selective advantage.

But much of the knowledge we acquire is not like this at all; it consists of arbitrary forms whose regularity or structure results simply from the fact that they occur often, or are repeatedly associated with reward and
Figure 1.2  (A) An array of 200 randomly positioned dots. (B) Each dot in the array of A has been paired at a position mirror-symmetric about the vertical midline. (C) Each dot has been paired at a position up and to the left of the original position. (D) Each dot has been paired at a position displaced radially and tangentially from the center. It is the structure that leaps to the eye, though this is technically a form of redundancy; the random positions of the dots contain much more information, but the eye gives them less prominence.

gratification. Each individual system has to discover these for itself, and we spend our lives finding, storing, and using knowledge of these regularities in our sensory diet. They range from the often-repeated experience of our parent's smell, voice, and appearance, through the geographic details of our environment and the acoustical specificities of our language, to the customs, myths, and true knowledge of our culture. Much of this process of acquisition is fostered by teaching, but each individual brain has to do a lot of discovering for itself.

So far we have been considering the goal of the computations the cortex performs on the current sensory input, arguing that it prepares a representation suitable for discovering associative structure, and that this process entails storing world knowledge. But each cortex not only has its own experience and history, but also an evolutionary history. Evolution results from natural selection acting on variants produced genetically, so perhaps
the pattern of variants produced by the cortex has enabled it to excel in the evolutionary acquisition of world knowledge.

Evolutionary Learning and Neocortex

Most people will accept the fact that there is such a thing as inherited knowledge of the world. Many of the most striking examples are found in insects—for example, the yucca moth could not fertilize the yucca plant and use its ovaries as incubators for its own eggs without such knowledge, nor could the ichneumon select a particular species of caterpillar to lay its eggs in. But it occurs in mammals too—the specialized skills of a retriever are quite different from those of a sheepdog or a greyhound—and no one doubts that these skills have a large inherited component.

Now a characteristic cannot play an important role in the evolution of a species unless it is controlled genetically and subject to genetic variability. Therefore the view that neocortex is responsible for our rapid evolution implies that its function must be controlled genetically, for otherwise it could not have brought us to the position we are in. The full hypothesis must therefore be that the neocortex gives us useful knowledge of the world in two ways: not only does it discover the structure of its world by experience during its lifetime, but it also has mechanisms, adapted through the process of genetic selection, that confer skills for doing this. These mechanisms and skills are sometimes highly specialized and amount to inherited knowledge of the world. On this view both extremes of thought about the origin of our mental powers are correct: the neocortex acquires knowledge of the world by nature as well as by nurture, but these methods work toward the same end rather than being the mutually exclusive alternatives that we tend to think. For this reason they can be considered together when trying to define the computational goal of the neocortex.

REPRESENTATIONS DESIGNED FOR KNOWLEDGE ACQUISITION

What we know of the neuropsychology of neocortex does not at first suggest that it is concerned with acquiring, storing, and utilizing knowledge of the world. Instead, it seems to form representations of the current scene in the sensory areas, and perhaps the motor area could be thought of as forming a representation of current motor actions. But this representational function does not necessarily conflict with the hypothesis about acquisition of knowledge. Different types of representation are suitable for different purposes, and the cortical representation may be one that is specially adapted to facilitate the learning of new associations. It turns out that storage and utilization of knowledge about the world is necessary to form such a representation, so the comparative anatomists' view that neocortex provides the filing cabinets of the central executive could be nicely reconciled with the neuropsychological facts about representation.
Figure 1.3 Flow diagram for perception suggested by the hypothesis that the cerebral cortex creates a representation of the current sensory scene that facilitates the identification of new associations. To separate new information from knowledge (i.e., redundancy) there must already be a store of known structure and regularities found in sensory inputs, and this must be used to form a model that accounts for as much as possible of the current sensory input; what this accounts for is then removed from the input representation. Though we think we experience sensory messages directly, what we see corresponds better to the contents of the heavily outlined boxes.

Figure 1.3 illustrates a flow diagram for perception according to this hypothesis. The sensory messages are combined with a store of knowledge of the world to find the best model of the current sensory scene. This model is then compared with the sensory messages being received, and those parts that match are removed. The residue represents the part of the current sensory input that is unaccounted for by preexisting knowledge. Ideally, this would correspond to new information, together with noise of random origin. We can be aware of this residue, but the subjectively salient and objectively useful parts of the sensory flow consist mainly of items that have been accounted for (i.e., items that have been successfully modeled), and also new regularities or structure in the parts that the current model does not account for. This flow diagram has features related to the “matching response” of MacKay (1955), the thalamic “active blackboard” of Munnford (1991, 1992), and the adjusting feedback of Daugman (1988) and Peco (1992).

Changing the Code Stores Associative Structure

The suggested operation can be thought of in a different way, as a recoding to reduce redundancy. The presence of one type of associative structure
in a body of data makes it more difficult to detect another type, so to
detect this new type it is desirable to recode the messages to eliminate
the first type. It is certainly very often the case that removing a known
type of associative structure makes it easier to identify a new type, and
figure 1.4 provides an illustration. The left part is a normal image and
thus has an autocorrelation function that extends over a large fraction of
the whole image. The counterpart of this is the great excess of low spatial
frequencies in the power spectrum of the Fourier transform, and these
can be removed by applying an inverse spatial filter to make the power
spectrum level. This process of “whitening” eliminates the correlations,
estimated over the whole image, between pairs of points with any fixed
separation, and the result is shown in the right image. It is clear that the
higher order structures, whatever they are, that correspond to borders and
edges survive and can be more easily examined in this image.

In outline then, the idea is that associative structure one already knows
about should be removed from the data stream to make it easier to detect
new associative structure. Knowledge of the old associations should be
used to change the code and thus modify the representation so that these
old associations are no longer present. This is the idea of recoding to
reduce redundancy (Barlow 1959; Watanabe 1960) or, if you like a simpler
analogy, it is like calculating the regression that corresponds to an already
recognized correlation to make it easier to find further relationships in the
residuals. Of course the modifications will not generally be as simple as
subtracting out an unexpected regression, but a set of modifications aimed at
accounting for and reducing the known structure in a set of images would constitute stored knowledge about those images.

To an outside observer, a system performing these operations would look like one that constructed Craik's working models (Craik 1943) and Tolman's cognitive maps of the environment (Tolman 1948), for it would show evidence of finding and using the associative structure that underlies such models and maps. Obviously this store of knowledge has many other potential uses, particularly in the processes of imagination and recall where we experiment and play with what we know. Possibly it could be made accessible in the absence of sensory input by lowering neural thresholds in the box marked "stored knowledge about the environment" in figure 1.3, but this possibility cannot be pursued here. Using this knowledge to discount the expected in the representation of the current scene would have enormous selective advantage by improving learning and the acquisition of new knowledge, though it certainly seems wasteful not to use it for imagination and recall as well.

Anything that improves the appropriateness and speed of learning must have immense competitive advantage, and the main point about this proposal is that it would explain the enormous selective advantage of the neocortex. Such an advantage, together with appropriate genetic variability, could in turn account for its rapid evolution and the subsequent growth of our species to its dominant position in the world.

Although these notions do not obviously follow from the neurophysiological facts, I think suggestive evidence in support can be found from the changes in neural connectivity that occur in the sensitive period early in the life of cats and monkeys (Hubel and Wiesel 1970;Movshon and Van Sluyters 1981), and in the known phenomena of pattern-selective adaptation discussed elsewhere (Barlow 1990, 1991). There are aspects of the evolution and neurophysiology of the cortex that we certainly do not yet understand properly and the new hypothesis can give us a fresh viewpoint if we examine what it requires in more detail.

Acquiring Knowledge from Representations of Features

Acquiring knowledge means finding out about the regularities and patterns in the sensory input. It's a vast task to determine the associational structure of the continuous stream of sensory messages that we receive, and table 1.1 lists some of the requirements, starting with the point above about the desirability of removing evidence for the associations you already know about. The next items have been discussed before (Barlow 1991) but will be summarized below.

Suppose that the representation of the current scene consists of reports of features, of which there can be a wide variety. For instance, one of them might be a point in the image having a luminance value above the mean for the neighborhood of that point, and this would correspond approximately to the feature that causes the firing of an on-center ganglion cell in the retina.
Table 3.1  What would make it easier to identify new associations?

| Remove evidence of the associations you already know about | To facilitate detecting new ones |
| Make available the probabilities of the features currently present | To determine chance expectations |
| Choose features that occur independently of each other in the normal environment | To determine chance expectations of combinations of them |
| Choose “suspicious coincidences” as features | To reduce redundancy and ensure appropriate generalization |

Or it might be the occurrence of a visual pattern resembling a monkey’s face, which would correspond to the occurrence of the trigger feature of a so-called face cell in inferotemporal cortex. Thus almost any representation one can imagine can be described as reporting the occurrence of features.

There must be many levels in the actual representational system in the brain, and more complex features are presumably represented at higher levels. The first item in table 3.1 suggests that recoding to take account of identified regularities in sensory messages will be an important step in progressing to higher levels in the perceptual system. But for present purposes let us consider a single level and examine what is needed to identify a new association. We hope that the repetition of this one operation may lead to a system that identifies the complex associations that we undoubtedly use all the time.

The Need for Prior Probabilities

To identify new associations a representation must do more than just report the occurrence of features: it must also signal the unexpectedness of the features reported, or at least make this information immediately accessible. This might be done by adjusting the threshold for a unit so that, averaged over a long period, it fires once in a particular period; when it fires, it then signals an event that has a probability of occurring once in that period. Alternatively, the number of impulses in the volley signaling an event might be an inverse function of its probability such as \(-\log p\). Either of these would appear as forms of habituation, which is of course often observed in sensory systems.

The reason prior probabilities are needed is obvious: to show that two features are associated one must show that they occur together at a rate different from that expected by chance, and to calculate this expected rate one needs to know the expected rates of the constituent individual features. Of course one also needs to know how often the features occur together, but one can justifiably regard this as a requirement of the associative mechanism itself, while it seems more natural to suppose that the representation is responsible for storage and access to the rates of the individual features.
Two points may need clarifying. First we are assuming that the probability of occurrence of a single feature can be estimated from its rate of occurrence over some period in the recent past. This would not always be justified, but in some cases it will be and to let the argument proceed let us assume it is. Second, the features we shall be dealing with will usually have prior probabilities well below half; this means that the predicted rate of occurrence of joint features will be low, and their expected number may be close to zero. Under these circumstances it becomes difficult to establish a negative association, and one must therefore look for joint features that occur more often than expected by chance. That is why they were called suspicious coincidences or clichés (Phillips et al. 1984; Barlow 1985), but the basic property is their nonaccidental nature.

The Need for Independence

Knowing and using the unexpectedness of features seem unavoidable for efficient associative learning, but there is another highly desirable property when detecting new associations, namely statistical independence, in the environment to which the system is adapted, of the features represented. Even in a simple case, such as finding a new association between a special occurrence such as reinforcement and an individual element of the representation, one would start by assuming they were independent to estimate the expected number of coincidences. The alternative would be to take account of the known correlations, but this would become difficult when detecting associations between arbitrary pairs of elements, and virtually impossible if one wished to find an association with some logical function of a group of elements. In that case one either has to know the associational structure within that group, or else one must again assume independence, and if one is going to do the latter it is important to make sure that the events represented are in fact as nearly independent as possible. While it is plausible for a representation to store the rate of occurrence of its individual elements, one cannot suppose that it stores the associational structure of arbitrary groups of such elements.

Are we actually able to detect new associations with logical functions of representational elements? For simple functions, surely we can, and so can most animals. We learn to stop at red traffic lights and not at green ones, for example. In this case one might suppose that there are different representational elements for red and green lights, but it would be a great restriction on the utility of a representation if this was always necessary before separate associations could be formed. Harris (1980) brought this out very nicely when discussing contingent adaptation, for he noted that almost any contingency that had ever been tested seemed to produce adaptive effects. How could this be, he said, if contingent adaptation requires neurons specifically sensitive to each contingency? We might have neurons signaling yellowness, and perhaps Volkwagens, but surely we cannot have neurons reserved for signaling yellow Volkwagens! This problem
will be considered again, but the advantage of distributed, as opposed to grandmother-cell, representations results from their supposed ability to utilize the vast number of combinations of active elements, and this advantage would vanish if, as a result of their prior probabilities being unavailable or grossly misleading, one could not form associations with these combinations efficiently.

Following the idea that one should discount the expected, a possible course of action would be to devise a code in which the elements occur as nearly as possible independently, and some ways of doing this have been suggested elsewhere (Barlow 1959, 1989, Barlow and Feldman 1989; Hentschel and Barlow 1991), together with evidence that something of the sort may be happening (Barlow 1990). As already pointed out, the codes that are required to obtain independence embody knowledge about the associational structure of the environment, and an outside observer watching behavior based on this modified representation should suspect that some kind of cognitive map or working model of the environment had been constructed.

Forms of Representation

A distributed representation is one in which the features that can be utilized effectively for further processing are represented by combinations of activity of the elements, rather than directly by the activity of neurons or elements specifically and selectively sensitive to each of these features. The 7-bit ASCII code provides a familiar example of a distributed representation, and because one must perform a logical manipulation on the representational elements before one can decide if the represented feature has occurred, we say that they represent the features implicitly or indirectly rather than directly. Tony Gardner-Medwin and I have been exploring a limitation on the use of implicit representations for learning (Gardner-Medwin and Barlow 1992, 1994).

The limitation arises as follows. Consider classical conditioning, where an initially neutral sensory feature is "reinforced" by being presented repeatedly in conjunction with a pleasant reward such as some food in the mouth. When the animal has identified the association, it uses the initially neutral feature to predict the reward. Now to determine whether there is a genuine association one must form a 2×2 contingency table for the feature and the reward, counting the numbers in each box of this table. If the feature is directly represented there is no great conceptual difficulty in obtaining all these numbers; assuming that knowledge of reinforcement is available everywhere, then local mechanisms at the element can estimate how often the feature and reinforcement occur by themselves, how often they occur together, and how often nothing occurs. A calculation equivalent to a chi-squared test can then be done on these numbers to decide if the association is genuine. In the case of implicitly represented features this is not so straightforward.
The difficulty is that there is no point in the system where all the information is available to estimate the necessary numbers. One can imagine the reinforcement signal being available at all the elements that carry the information telling one that the feature has occurred, but one of these elements by itself is not enough to determine whether that feature occurred or not. One must evaluate the logical function using all the elements before one knows this. One can postulate an element that does this logical decoding, but such an element would directly represent the feature and it would no longer be only implicitly represented. What can be done?

Even though accurate counts of the required numbers are not available at any one spot, a relatively simple mechanism could collect together information for an inaccurate estimate. In the logical representation of the feature there will be some elements that are positively correlated with the presence of the feature, and others that are negatively correlated. The appropriately weighted sum of the activities of these elements will give an indication of the presence of the feature, and the average over time of this measure can be used to estimate how often it occurs. We know from the limitations of perceptrons (Minsky and Papert 1969) that there are many logical functions that this method will be incapable of detecting correctly, but we also know there are many cases where it works satisfactorily. An approximation of this sort actually seems to be involved in most learning in artificial neural networks. We have been trying to determine for what types of representation such learning will be reasonably fast and efficient, and under what conditions it is bound to be slow and unreliable.

Statistical Efficiency as a Measure of Performance

To assess the merit of one representation against another, we need a measure of associational performance, and we have used the statistical efficiency defined by R. A. Fisher (1925) for this purpose. To make any statistical decision up to a required standard of reliability a sample of a certain minimum size is necessary, but if the method is inefficient a larger sample will be needed to obtain the same standard of reliability. Fisher's efficiency is simply the ratio of these two sample sizes. In our case the sample size is given by the maximum number of occurrences or coincidences that could occur in the time for which the counts were made, so if the method is inefficient it will take longer to determine that an association is present, or more mistakes will be made if a decision is made in the same time. It is pretty clear that speedy and reliable learning about new causative factors in the environment will have high survival value, and the statistical efficiency attainable in a particular type of representation gives a very direct measure of how useful it would be for enabling an animal to detect new associations and acquire new knowledge of the world.
Explicit Representation

So far we have referred to "directly represented features," where there are selectively sensitive representational elements that respond when and only when the feature is present and "implicitly represented features," for which there are no such selectively sensitive elements but the appropriate logical analysis of the pattern of activity in the whole representation nonetheless allows one to decide if the feature is present. We now introduce an intermediate type, explicitly represented features, for which we have been able to solve the problem of determining the statistical efficiency for detecting associations.

An explicitly represented feature is one whose presence can be determined by a simple logical operation performed on a subset of the elements in the representation, rather than on the whole representation. So far the simple logical operation we have analyzed is the presence of a particular pattern in the elements of the specified subset, since this seems both the simplest and perhaps the most interesting case. It turns out that inactive elements carry very little information if the representation is reasonably sparse (i.e., the average proportion of elements active at any one time is low, say less than 10%), so one need consider only the active elements. Each of these directly represents a different feature, so the occurrence of the pattern corresponds to the conjunction of joint occurrence of certain specific features. To return to Harris's example, if one element at a particular point in the visual field directly represented Volkswagens, and another element at that position directly represented Yellowness, Yellow Volkswagen would not be directly represented, but they would be explicitly represented by the joint occurrence of the above two elements. The question we think we have answered is "Under what conditions can a representation in which there are Yellowness (Y) units and Volkswagen (V) units, but no Yellow Volkswagen (YV) units, nonetheless be used to detect efficiently an association with Yellow Volkswagens?" The answer is, however, more general than this, for it applies to multiple conjunctions and patterns in subsets with more than a pair of active elements.

The outline of the analysis is as follows. To determine if an association is present between a feature and reinforcement (R) one does a chi-squared test on a 2×2 contingency table in which the feature (Y, V, or YV) is one of the variables and reinforcement (R) is the other. Because of the sampling errors in the numbers in such a table the result will be variable, and this variability determines how large a sample is required before one can confidently assert that an association is present.

If there are no YV units one must look at the 2×2 table for Y vs. R and V vs. R, and combine the results to assess whether YV is associated with R. Now for each of the Y vs. R and V vs. R tables there is a perturbing factor: Yellowness can occur with reinforcement even if there is no Yellow Volkswagen present, and likewise for nonyellow Volkswagens. These intrusive extra occurrences will not bias the result if one knows the unexpectedness
of $Y$ and $V$, but they will add to the variability of the two subtables, and even when optimally combined the decision about the association of reinforcement with Yellow Volkswagens cannot be made as reliably as if they were represented directly.

**Sparse Coding Helps**

How serious is this factor? Note first that the problem arises when the same representational element is active in more than one of the features that may be reinforced. In the current example, the $Y$ unit is active for all yellow things, not just yellow Volkswagens, and similarly for the $V$ unit. The extent of this overlap depends on the sparseness of the representation, which is defined by the average fraction of the elements that are active. If it is very sparse, then only a small proportion of the units will be active for any input, and there will be little overlap. Indeed, if it is sparse enough there will be only a single unit active for each input, and each will therefore be directly represented. On the other hand if it is dense, a given unit will be active for a high proportion of inputs, and the overlap problem will be serious.

When there is overlap, what matters is the number of extraneous extra occurrences relative to the number of genuine occurrences of Yellow Volkswagens, and this in turn depends upon the probability of the joint event ($YV$) relative to the single features ($Y$ and $V$). If Yellowness and Volkswagens are both common, then the reinforcement of rare Yellow Volkswagens would be masked by the quite frequent chance reinforcement of other yellow things and other colored Volkswagens.

One can show that the efficiency for detecting a feature $X$ depends to a good approximation upon the value of a parameter $\Gamma$, that is equal to $\alpha_p Z/K(\alpha)$, where $\alpha_p$ is the fraction of representative elements in the subset active for the representation of the feature $X$, $\alpha$ the average fraction active for all inputs, $p$ is the probability of the feature $X$, and $Z$ is the number of representative elements in the subset. Figure 1.5 shows how efficiency varies with the value of this parameter.

**Improbable Features Need Denser Representation**

Note first that efficiency increases with $\Gamma$, and, as expected, $\Gamma$ increases with the number of neurons in the subset and the average sparseness (low $\alpha$). It also increases both with the probability $p$, of the feature $X$, and with the activity ratio $\alpha_f$ for the feature $X$; in fact these two are reciprocally related, so a very rare feature $X$ can still form associations efficiently if it causes an unusually large proportion of the units in a subset of the representation to be active. Clearly there is scope here for genetic factors to improve selectively the performance of a learning network: factors of biological importance should cause many units in a learning network to become active. Another way of putting this is to say that many of the
The curve shows how statistical efficiency for detecting associations with a feature \( X \) varies with the value of a parameter defined as follows: \( \Gamma = \alpha p_x Z / \bar{\alpha} \), where \( \alpha_x, \bar{\alpha} \) are the activity ratio for feature \( X \) and the average activity ratio, \( p_x \) is the probability of \( X \), and \( Z \) is the number of neurons in the subset under consideration. For instance, one could identify an association with any one of the 45 possible pairs of active neurons in a subset of 10 with an efficiency of 50% provided that the neurons were active independently, the pair caused two neurons to be active, the probability of the pair occurring was 0.1, and the average fraction active was 0.2. (From Gardner-Medwin and Barlow 1994)

directly represented features should correspond to features possessed by biologically important objects; then, when one of these objects appears, it will cause a high level of activity.

Next look at the actual efficiencies attainable for various values of \( \Gamma \). Although one needs \( \Gamma \) to be 10 or 100 for efficiencies in the 90% to 100% range, useful efficiencies of about 50% are obtained with \( \Gamma = \sim 1 \); this is the order of magnitude of the efficiency of human subjects detecting bilateral symmetry in dot patterns such as those shown in figure 1.2 (Barlow and Reeves 1979). Consider a subset of 10 elements in a network; if one could specify 10 mutually exclusive features, the elements of the subset could each handle one of them and associations with them could be formed with 100% efficiency. Now suppose that the features of interest do not all cause firing of only single elements among the ten. If a particular feature \( X \) does cause firing of just one element (\( \alpha_x = 0.1 \)) but this element is also active in conjunction with other elements when other features occur, then if \( p_x = \langle \alpha \rangle = 0.1 \) we will have \( \Gamma = 1 \) and the efficiency for detecting associations will have dropped to about 50%. This reduction occurs because intrusive or accidental reinforcement occurs in conjunction with the activity of any given element, but this is a small price to pay for the increased versatility resulting from the possibility of using and learning associations of combinations of the features, as illustrated below.

Suppose the feature \( X \) is represented by the conjunction of two elements (\( \alpha_x = 0.2 \)). If again \( p_x = 0.1 \) but we suppose (\( \alpha \)) is now 0.2, the same as the new value of \( \alpha_x \), then \( \Gamma \) still has the value 1, corresponding to the same efficiency, ca. 50%. There are 45 such conjunctions of pairs of elements among 10 elements, so a much wider range of features can be used to form...
associations efficiently, and there is not an enormous loss of efficiency compared with the direct representation of features on single elements. Notice that the above applies to features represented by pairs of active units, but a particular merit of such a system is that it can form associations with patterns containing three or more active elements. Even if such multiple conjunctions of directly represented features are rare, provided that they cause activity in a high proportion of elements, they will be learned with reasonable efficiency.

What this shows is that it is possible to learn about explicit conjunctions of any number of elements in known subsets of a representation, provided that the representation is sparse, provided that these conjunctions do not occur too infrequently and activate a substantial proportion of elements when they do, and provided that the representative elements can be considered, a priori, to occur independently. How to achieve this, and read out the results in a useful way, cannot be gone into here.

The analysis we have done so far is only a beginning. What can be done using the union rather than conjunction of representational elements in a subset? What can be done with threshold logic functions on the activity of members of a subset? We do not know the answer to these questions, but one point does seem evident.

We have already seen that the features that are directly represented should (1) occur as closely as possible independently of each other in the environment to which the representative system is adapted; (2) occur sufficiently frequently so that the representation is neither too dense nor too sparse. These requirements might not be too difficult to meet if one could postulate an indefinite number of directly represented features, but such an indefinitely large representation would have none of the capacity to generalize sensibly that is needed in a representation to be used for learning. This introduces another requirement for the selection of directly represented features: they must each represent as much as possible of the incoming stream of data from the environment and must occur frequently so that generalization occurs usefully. Some tests of this prediction on digitized images of natural scenes will now be described.

**SELECTION OF DIRECTLY REPRESENTED FEATURES**

It is generally agreed that the neurons of the primary visual cortex respond selectively to the borders and edges of objects in the visual image. There is no argument about whether they should be regarded as edge-detectors, Gabor filters, or wavelet functions, but there is no disagreement that they do in fact respond to the oriented patterns of light that occur at the borders of objects. If the arguments (Barlow 1985) about the importance of nonchance associations are correct, then measurements of the distribution of light at the borders of objects should show that edges qualify as "suspicions coincidences." We set out to test whether this was so, and the main result confirms that it is (Barlow and Tolhurst 1992).
We took a selection of digitized images and removed the correlations between pairs of points, averaged over the whole image, by the "whitening" process described before and shown in figure 1.4; this leaves behind the image structures we are now interested in that occur at the borders of objects. The distribution of pixel values in such whitened images gives us the basis for the chance expectation of combinations of pixel values, and what the hypothesis says is that at the borders of objects we shall find combinations of pixel values that occur more frequently than this chance expectation.

Perhaps it is already obvious by inspection of the whitened figure that this is the case, for you would not expect to find by chance the rows of high or low values you can see in figure 1.4. To confirm this we measured the distribution of the sum of nine pixel values selected at random from all over the whitened image to provide the chance distribution, and from nine adjacent spots in a row to show what actually occurs. Figure 1.6 shows the result: the distributions are strikingly different. For the sum of nine randomly selected pixels the range is from about 980 to 1320 on the horizontal scale, but as you can see values outside this range are very common for the sum of nine pixels in a row.

Do these extreme values occur at the borders? Yes they do, as shown in figure 1.7, which marks the positions in the image where these extremes occur. As you see they occur at edges.

Is this a consistent feature of the sum of pixels in a line? To answer this we looked at a varied selection of 15 images, and estimated the kurtosis excess for the sums in a line compared with randomly selected pixels and sums over square regions. This measure (Weatherburn 1961) is based on the fourth moment and values greater than 0 can be crudely taken to indicate that the distribution has an excess of extreme values compared with a gaussian. As shown in table 1.2, the kurtosis excess is much greater for the line sum than for the other distributions, though it has to be said that we do not understand why patches, and even single pixels, also show kurtosis excess. The large excess for lines, combined with the fact that the extreme values occur at the borders, vindicates the hypothesis that the features we use to represent an image are suspicious coincidences—at least in the case of the orientationally selective units of V1.

**SUMMARY AND CONCLUSIONS**

It was suggested initially that we dominate the world because we know more about it than other animals, and that it is the neocortex that is responsible for this. How to acquire and store knowledge of the world is a vast problem, but although we have only scratched the surface we may be beginning to discover how the neocortex could, as the combined result of genetic selection and individual experience, provide us with a representation of the current scene that automatically stores, gives access to, and adds to such knowledge.
The first principle, suggested diagrammatically in figure 1.3, is that neocortex removes associative structure that has already been identified through past experience. This is analogous to discounting the mean luminance in light adaptation, or removing a known regression when trying to make sense of residual deviations. Identified structure would be stored, and when recognized in the current scene it would form part of a matching model; the unmatched residue would contain new information about that scene. Stored knowledge of the associative structure of the world would be used continuously and automatically in this way, but there could be other methods of accessing it for purposes of imagery and recall.
Figure 1.7 The left panel is the whitened image from figure 1.4. In the right panel white dots are placed at the positions of the upper extreme values of the distribution of the sum of pixels in a line shown in figure 1.6, and dark dots at lower extreme values. The extreme values obviously occur at the borders of objects in the image. Hence the combinations of pixel values that occur at edges are ones that would not be expected by chance.

Table 1.2 Kurtosis excess for 15 images

<table>
<thead>
<tr>
<th>Description</th>
<th>Mean ± SE (n = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of pixels over line 9 x 1</td>
<td>15.82 ± 3.46</td>
</tr>
<tr>
<td>Sum of 9 random pixels</td>
<td>0.699 ± 0.76</td>
</tr>
<tr>
<td>Sum over square 3 x 3</td>
<td>7.26 ± 1.68</td>
</tr>
<tr>
<td>Single pixels</td>
<td>8.81 ± 1.70</td>
</tr>
</tbody>
</table>

2. To prove that an association between two features exists you need to know their individual frequencies of occurrence, because you must estimate the chance frequency of joint occurrence to show that the actual frequency is significantly greater.

3. To detect associations with combinations of features, the features should be chosen so that they occur as nearly as possible independently of each other in the environment to which the system is adapted, for otherwise the expected frequency of occurrence of a combination is hard to determine.

4. In distributed representations, detecting associations with conjunctions of features is difficult because accidental associations with the constituents of the conjunctions mask associations with the conjunctions themselves. This problem tends to make the identification of associations with conjunc-
tions inefficient in dense distributed representations, and such representa-
tions are therefore unlikely to be useful.

5. On the other hand in sparse distributed representations associations can
be identified reasonably efficiently (say 50%) with features represented by
the conjunction of directly represented features, provided that these con-
junctions are not too sparsely represented and occur with a frequency not
too far below that of the directly represented features.

6. To generate a reasonably economical representation of the current scene
the directly represented features should be suspicious coincidences—com-
binations of signals from lower levels that occur frequently but would be
rarely expected by chance. The representational elements at higher levels
should be matched to the biological importance and statistical structure of
occurrences at lower levels.

7. This notion that the representational elements used by the brain corre-
spond to suspicious coincidences, or combinations of simpler events that
occur more often than expected by chance, has received some support from
the statistical analysis of edges in images.

8. The features that are directly represented at any level in the hierarchy
will have a strong effect on the performance of a representational network,
including the way that generalization occurs. The selection of these fea-
tures is likely to be one way that genetic factors exert their influence. In
addition, ontogenetic control of the connections between levels probably
determines the way that information of different types is segregated and
brought together according to nontopographic principles. In these two
ways, and possibly others, genetic factors must influence how the cortex
handles sensory information, and they can be regarded as an inherited
store of world knowledge; the genetic variability that has enabled such a
store to be formed may be at least as important as the ability of the cortex
to acquire world knowledge by its own direct experience.