Biol. Cybern. 66, 241–251 (1992)

Biological Cybernetics

On the computational architecture of the neocortex

II The role of cortico-cortical loops

D. Mumford

Mathematics Department, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Received June 29, 1991/Accepted July 12, 1991

Abstract. This paper is a sequel to an earlier paper which proposed an active role for the thalamus, integrating multiple hypotheses formed in the cortex via the thalamo-cortical loop. In this paper, I put forward a hypothesis on the role of the reciprocal, topographic pathways between two cortical areas, one often a 'higher' area dealing with more abstract information about the world, the other 'lower', dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops. Evidence for this theory is reviewed and experimental tests are proposed. A third part of this paper will deal with extensions of these ideas to the frontal lobe.

1 Introduction

The point of view of these papers and the motivation behind them was described in the introduction to the first part of this paper. Summarizing, the idea is that the uniformity and highly specific layered structure of the neocortex of mammals suggests that some quite universal computational ideas are embodied by this architecture. This paper presents two proposals for computational mechanisms embodied in this structure.

The first part dealt with a conjecture for the role of the reciprocal and largely topographic pathways connecting each area of the cortex with a corresponding nucleus in the thalamus. It was proposed that a very important part of the computation performed by the cortex made use of this loop:

- that the cortex learns multiple patterns that recur in sensory stimuli and in stereotyped motor output,
- that at any given time, the cortex is attempting to analyze the present situation in terms of these patterns and, in so doing, generates multiple hypotheses, often conflicting,
- that all these hypotheses are sent down to the thalamus where a kind of voting takes place in the dendritic arbors of the thalamic neurons,
- that the consensus is then broadcast back to the cortex as an updated view of that aspect of the world dealt with by that area of cortex.

This theory was summarized by describing the role of the thalamus as that of an 'active blackboard' bearing the data on which the cortex was working.

The second part will deal with the reciprocal and largely topographic pathways that are found throughout the cortex connecting pairs of cortical areas. A detailed proposal will be made for the nature of computation performed by exchange of messages via this loop. The present paper will expand these ideas for sensory processing carried out in the posterior half of the cortex and a third part of the paper will apply them to the computations underlying planning and action carried out in the frontal lobe. I will propose specific tests for some of these ideas. As in the earlier paper, we have included a good deal of background both on neuroanatomy and on computer science in order to make the ideas as clear as possible to readers from various specialties. Finally many people have had ideas similar in various ways with ideas presented below: the ones I know of are the 'Adaptive resonance theory' of Carpenter and Grossberg (1987), the 'HyperBF' theory of Poggio and collaborators (Poggio 1990), the 'countercurrent' processing theory of Deacon (1988) and recent theories of Rolls (1990) on the 'back-projections' in the brain.

2 Pyramidal neurons and cortico-cortical pathways

I begin by reviewing some neuroanatomical facts about cortical pathways. As described above in the first part of this paper, each hemisphere of the primate cortex seems to be divided into something of the order of a hundred areas each with a specialized role. There are, of course, species differences[1], but the general map and often many of its details are roughly homologous for most species.

Tracing pathways supplements the map of cortical areas with a diagram of their interconnections. These interconnections turn out to be relatively sparse, in the sense that of the 10,000 possible one-way pathways that could exist between 100 areas in each hemisphere, perhaps only the order of magnitude of 2000 exist (Felleman and Van Essen 1991). This could well be the result of limitations of space inside the cerebral hemispheres, which can only contain so much white matter, but it obviously has computational significance. A very important fact, central to the theory in this paper, is that all or almost all (there are some ambiguous cases) interconnections so far discovered are reciprocal: *if area A projects to area B, then B projects to A.*

What types of cells set up these pathways? There are two main types of neuron in the neocortex: pyramidal cells and interneurons. Pyramidal cells are large, excitatory, with a pyramid shaped cell body, spiny dendrites and a long myelinated axon (myelin is nature's way of insulating axons to ensure stronger, faster long-distance signals) that projects to another area of the brain (another cortical area or subcortically), usually with branches projecting locally[2]. These are the neurons which create these cortico-cortical pathways. Interneurons are small with only local projections, spineless dendrites and are usually inhibitory. An intermediate class consists of the spiny stellate cells populating layer IV which generally project only locally but resemble pyramidal cells in being excitatory and having spines: they are roughly pyramidal cells without a long axonal projection, and are sometimes called small pyramidal cells.

The percentage of neurons in human cortex which are pyramidal has been variously estimated as 60%–80%, although there seems to be some doubt about counting accurately the interneurons which have smaller cell bodies[3]. This distinguishes the structure of the cortex strikingly from other bodies such as the olfactory bulb and the cerebellum, in which interneurons substantially outnumber the cells with long axons.

Some crude numerical estimates may be useful: using the estimates cited in the first part, each hemisphere of the human cerebal cortex may contain roughly 10 billion neurons. Then an average area would have about 100 million neurons, with say 60 million pyramidal cells projecting to some other cortical area. If this area is connected to 30 others, each pathway comes out as containing the order of 2 million fibres, the same order of magnitude as the optic nerve. In strong contrast, the cerebellum has an order of magnitude more neurons than the cortex, and most of these are its principal interneurons, the granular cells, whose number is estimated at *100 billion* in man! The number of Purkinje cells, its output cells, is only 15 million or .03% of the total (Ito 1984). The pathway between the cerebellum and the rest of the brain is also an order of magnitude bigger than the cortical pathways: in man it is estimated to contain 20 million axons (Brodel 1981, p. 297).

The fact that the majority of cortical cells have inter-area projections, as opposed to exclusively intra-area projections, seems already to bear an important computational message: it means that almost nothing goes on internally in one area without this activity being transmitted to at least one other area. The classical view on the significance of the different areas of the brain was that it was similar to the modular decomposition of a computer program into subroutines. In computer programs, each module performs a specific task and the various modules pass input and output back and forth by means of messages (or via globally accessible data structures, like blackboards). The analogy suggests that each area in the brain has a specific capability, e.g. looking up words in a lexicon, sequencing motor acts, etc. and that the inter-area pathways exchange requests and answers. But this analogy would only make sense if the number of intra-area locally projecting neurons were an order of magnitude larger than the number of inter-area globally projecting neurons.

Since this isn't the case, a different paradigm must be sought. This is the main idea of this paper, which I will develop in stages. In essence, I want to propose that *the bulk of the computational work of the cortex is not carried out by one area at a time, but by information going back and forth over reciprocal pathways connecting pairs of areas: in doing this, each such pair of areas is trying to reconcile their constructs by some kind of relaxation algorithm.* Before developing this idea further, we need some more anatomical facts.

3 Higher versus lower areas

For a long time, there have been attempts, using the above mentioned modular view of the brain, to give each of the different cortical areas a particular functional significance and to describe the nature of the information represented by neuronal activity in each

[1] The major difference is that the number of areas increases with increasing brain size. For instance, the frontal lobe in humans is much larger than in other primates and seems to contain many more areas

[2] The existence of extensive *local* collaterals is a major difference between the output, pyramidal cells of the cortex and the output cells of the thalamus. It allows the cortex to carry on local calculations indefinitely without further stimulation, whereas the thalamus cannot do this

[3] Various references can be found in the discussion in DeFelipe and Jones (1988), pp. 590–599. The counts in Winfield et al. (1980) seem representative and are confirmed by immunochemical determination of the percentage of GABA cells. They find on average 67% pyramidal, 5% large stellate and 28% smaller interneurons (presumably inhibitory) in cat and rat cortex

cortical area. From such assignments, we can describe the pathways between the areas in terms of passing data from an area with one sort of concern to another. A persistent theme is to distinguish lower cortical areas, with direct sensory or motor connections from higher ones which are associating information from lower areas, so that information moves first from lower, more sensory areas to higher, more cognitive association areas and secondly from these association areas back down to lower motor areas.

There are several ways of establishing such functional correlations: firstly, the distance of an area from the nearest area with direct sensory or motor connections (the primary sensory and motor areas) is one indicator of how high-level it is. This is confirmed by comparative neuroanatomy, in that lower mammals have almost all their cortex taken up by the primary motor and sensory areas[4], while an increasing amount of secondary tissue appears in mammals with greater intelligence. Secondly, direct stimulation of the cortex of humans, first employed in operations for intractable epilepsy by Penfield, resulted in the patient's experiencing a variety of thoughts, ranging from very concrete sensations or motor reactions to quite elaborate memories or abstract ideas. Thirdly, the loss of functions in strokes can be correlated to the cortical area destroyed by the stroke. Fourthly, single cell recordings in animals, especially primates, enable one to correlate the firing of a particular neuron to the presence of various stimuli, or the performance of various tasks, and these show a clear gradient from elementary sensory or motor responses, to elaborate complex responses (e.g. the presence of a monkey's face in the field of view).

These four techniques give a fairly consistent, though imprecise, idea of which areas were 'higher' and which 'lower' than others and roughly what sort of data was being dealt with. Traditionally, one way in which this data has been put together is via a division of the cortex into primary sensory and motor areas, secondary sensory and motor areas and tertiary 'association' areas.

A much more precise way of ordering cortical areas, which agrees with and extends the above higher/lower ordering was found by analyzing the connections of the areas in terms of the layers of origin and the layers of termination of each pathway. To describe this, I need to first sketch the cell populations of the six layers. The pyramidal cells occur in two populations: the deep pyramidal cells in layers V and VI and the superficial ones in layers II and III. Layer IV in the middle is occupied mainly by the spiny stellate cells and, as we have seen, is the principal input layer for sensory data and other thalamic projections driving cortical calculation (although note that the cells in layer IV, by virtue of their arborization, will already perform some transformation on the input data). Layer I, called by Cajal the plexiform layer, has extremely few cell bodies of any kind, but is the zone for a rich set of connections between a second type of axonal input, the interneurons and the apical dendrites of the pyramidals. The inhibitory interneurons occur in all layers except I, and themselves break up into a dozen types or so with differing geometry and distributions.

In terms of layers of origin and termination, there seem to be three types of long distance cortex to cortex connections between areas, all set up by pyramidal cells. In this division, I am quoting the results in the exhaustive survey paper Felleman and Van Essen (1991). To see a particular example in detail, the projections to and from V1 are shown in detail in Perkel et al. (1986) and Van Essen et al. (1986). The survey paper deals primarily with the posterior, sensory-oriented areas of the cortex, and I will restrict the discussion at first to these areas. The first of these types of pathway originates in deep pyramidal cells, usually in layer V, and terminates heavily in layers I and VI, avoiding layer IV completely. The second of these originates in superficial pyramidal cells and terminates primarily in layer IV. The third of these originates in superficial pyramidal cells, but instead terminates outside layer IV, mostly in layers I and VI. There are a few reports suggesting further types of connections, but these, if present, don't seem to be widespread.

What makes this division into types impressive is that, whenever two areas *A* and *B* are reciprocally connected, and the previously discussed evidence shows clearly that if area *A* is 'higher', *B* is 'lower', in terms of their function, then:

1. *The ascending pathways* from *B* to *A* is set up by superficial pyramidals in *B* terminating in layer IV of *A*. Note that this is consistent with layer IV being the standard input layer at each stage of the stream of data all the way from the senses themselves to the highest cognitive areas.
2. *The descending pathway* from *A* to *B* always includes deep pyramidal cells in layer V of *A* terminating mainly in layers I and VI of *B*. If *A* is 'much higher' (in some loose sense, see (Felleman and Van Essen 1991)) than B, this is the only projection from *A* to *B*. Note that the projections from *A* to the thalamus are also set up by deep pyramidals (chiefly in layer VI), so we have a consistent picture of deep pyramidals projecting to lower structures, either in the cortex or sub-cortical. This is also consistent with the idea that *A* delivers a different kind of input to *B* from its standard input. I'll call these the *standard descending paths.*
3. *The descending pathway* from *A* to *B* may also include superficial pyramidal cells of *A* terminating chiefly in layers I and VI of *B*. This occurs if the ordering between two areas is not so clear, *A* is only slightly higher than *B* in Felleman and Van Essen's sense. Note that again terminations in layer IV are avoided by the 'higher' to 'lower' projections. I'll call these the *extra descending pathways.*

A summary of this pattern is shown in Fig. 1.

It would be nice if we could extend this picture unequivocally to the frontal lobe. While there have

[4] This is ignoring the limbic areas, dealing with emotion, social behavior and memory, which are also large in lower mammals

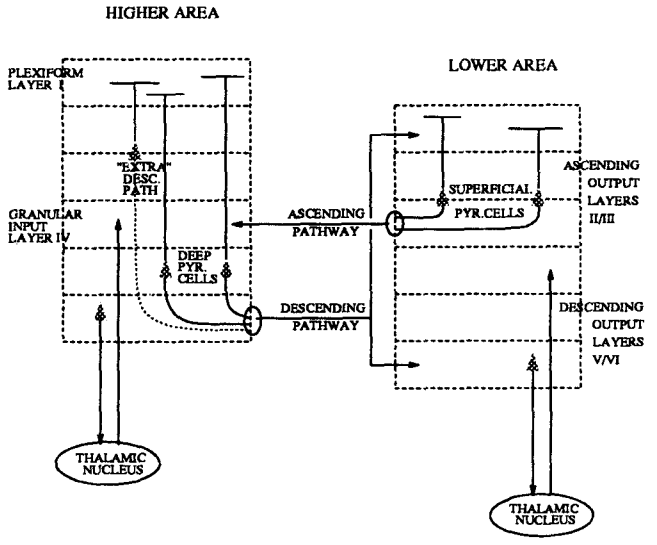

Fig. 1. Cortico-cortical pathways by layer

Fig. 2. Top down processing reveals the dalmation dog (Rock 1984)

been fewer studies of the laminar connections to and within the frontal lobe, present evidence seems to favor the idea that this same laminar pattern is present (Deacon in preparation; Primrose and Strick 1985). Felleman and Van Essen (1991), however, qualify this conclusion with the remark that "The patterns illustrated in the literature are difficult to interpret unambiguously . . .", (cf. section entitled "Hierarchical relationships in other areas", subsection "Somatosensory and motor cortex"). The data suggests a picture in which (*a*) the primary motor area, Area 4, is lowest, and the other frontal areas starting with the premotor and supplementary motor areas get higher and higher, in a complex pattern, while (*b*) the layers of the connections conform to the three types (*i*), (*ii*) and (*iii*) above.

Another important caveat is that when two areas *A* and *B* are reciprocally connected, one needn't be higher, the other lower, in any clear way. In this case, one can imagine that all types of connection are possible, and the neuroanotomy doesn't reveal anything directly about the functional nature of the pathway.

Can we make some sort of hypothesis about the functional role of the ascending and descending pathways? In the rest of this paper I will try to the analyze the role of these pathways in the sensory half of the brain, the occipital, parietal and temporal lobes, and leave to part III an extension of our theory to the motor half of the brain, the frontal lobe. Now the ascending pathways have never seemed to be problematic, because information must obviously flow from the senses up to cognitive areas. This ascending stream of information is referred to as 'bottom-up' processing. However, there is a general realization of the importance of 'top-down' processing too, involving the active use of high-level knowledge to help disambiguate low-level perceptions (as in the ability to discern the dalmation dog in Gregory's famous picture, see Fig. 2). This is what I want to analyse first.

4 Descending pathways carry templates

Let us step back and make some elementary observations about what descending pathways must do. Without any preconceptions about the computational nature of the cortex, one would expect that activity in lower sensory areas of the brain is directly correlated with elementary properties of the sensory input, while activity in higher areas is correlated with the presence or absence of some more subtle properties of the sensory input, e.g. the presence of a face. This is clearly born out by single cell recordings for instance (cf. Desimone's survey paper (1991) for a history and description of the so-called 'face' cells in inferior temporal cortex). One might say that the higher area is speaking a more sophisticated, more abstract language. In that case, when information is passed from a higher to a lower level of the brain, it must be translated from the abstract language of the higher area to the concrete terms employed by the lower area. If some particular pattern of bits in a higher area happens to mean 'face', there is no point sending this encoded pattern down to a lower area which knows nothing about faces. You have to translate 'face' into a signal in the terms used by the lower area, e.g. a pattern of bits signifying the appropriate concrete configuration of lines, shapes and colors. Such a translation is what in psychology would be called a mental image, a reconstruction of a detailed sensory signal that instantiates an abstract class of signals. In the language of pattern recognition, it is what is called a template. Early work in pattern recognition centered around the idea of recognizing classes of signals by having a specific template (i.e. a standard example of a signal in each class), which could be matched, feature by feature, second by second, or pixel by pixel, against the signal to be classified.

Our proposal is that the axons of the deep pyramidal cells in the descending pathways store templates in the weights of their synapses in the lower area. The single bit represented by a pulse on the axon of such a cell must stimulate, via the weights on its synapses in the lower area, a low level template-like response repre-

senting the translation of the information in that bit in the high level representation scheme into more concrete information in the low level representation scheme. One must not oversimplify here: a simplistic form of our hypothesis would be that specific deep pyramidal neurons, or small sets of them, were responsible for each template in the lower area. For example, one might suppose that several dozen deep pyramidals in inferior temporal cortex constructed an eye template, in the sense that the strength of their synapses in lower order visual areas created excitation equivalent to a retinotopic eye-like stimulus. This would be an elegant hypothesis, but it looks totally unbiological. Much more likely, it seems, is that the computation is distributed, that thousands of neurons are simultaneously carrying eye, nose, mouth, face, etc. templates. This would explain why recordings from such a large percentage of inferior temporal neurons show responses to faces, and that the ability to recognize faces is robust in the face of small local damage to cortex: all the usual arguments in favor of distributed representations in neural nets.

Some evidence for this hypothesis comes from the experimental fact that the axonal arbors of descending pathways are, on the whole, more extensive than those of ascending pathways: we would expect this if the descending pathways were recreating the pattern of excitation characteristic of some higher level construct, because such higher level constructs embody common extended patterns of excitation in the lower area. Very carefully drawn illustrations of the arborizations of typical V2 $\Rightarrow$ V1 axons can be found in (Rockland and Virga 1989). Her pictures suggest not only that a rather intricate excitation pattern is created by the top-down activity of such a neuron, but even that this pattern may reproduce the effect of extended lines, which would usually be part of any higher level visual pattern. This is seen in the fact that in some of her pictures the axon sprouts synapses at regular intervals, interspaced with synapse-free zones, as it extends in a specific direction in layer I; because of the known division of V1 into orientation-specific columns, this could well be the structure needed to stimulate successive parts of an extended line with a fixed orientation.

Recall that there are also extra descending paths formed by superficial pyramidal neurons, but only in case of areas which are not too far apart, one not being too much higher than the other. Our hypothesis is that the standard descending pathways which are always present carry templates, especially because the further apart two areas are, the most different will be their 'languages', hence the greater the need for template-like translations from one language to another. After I extend my hypothesis to the ascending pathways, I will come back and make some speculation on the role of these extra descending pathways.

5 Templates must be flexible

Early pattern recognition work using templates was never very successful, however. The difficulty was how to account for the range of variation in the objects being recognized: two eyes are never the same, and one must be able to recognize as eyes all the variations which normally occur, including eyes in people never seen before, eyes in strangely lit faces, cartoon eyes, etc. The problem of allowing for normal variations arises already when trying to classify some object on the basis of a few measurements: a classic example in the statistical literature was that of discriminating three species of Iris from the length and width of its petals. One must model the allowable variations of length and width within each species or, better, the full probability distribution of the measured features for each class before making an informed decision on the species from these two features. The problem gets harder for 1D signals such as speech: an essential adjustment is called time-warping, in which templates for the various phonemes are scaled to allow for the speaker's rate of speech. In 2D signals such as vision, the problem is much more difficult. Letters can be varied in many non-linear ways while still being readable, faces distort with differing expressions and shadows change the appearance of even simple industrial parts on an assembly belt. Other types of variation are not usually continuous but correspond to the object belonging to one of several subcategories: e.g. faces with and without glasses, a person being male or female, a screwdriver being plain or Phillips, etc. What all this suggests is that you need a flexible template: a template with built-in variability embodied in a set of parameters, whose values can be chosen so that the template will nearly match the example of the class in the signal being analyzed. In vision, early work in this direction is due to Fischler and Elschlager (1973), and a recent version can be found in Yuille (1991).

The parameters in a flexible template are not imagined to vary arbitrarily, but to have some restrictions placed on them which form an essential part of the template:

1. they generally have an allowable range, individually or jointly (e.g. two parameters may have to lie in some subset of the plane),
2. there may be a prior joint probability distribution,
3. one may store a set of useful examples – e.g. the parameter values for a prototype instance and some key borderline cases, showing the worst instances you've encountered.

How do these parameters and their allowable range fit into our neural theory? What I want to propose is that when a high level area has neurons which fire in the presence of eyes, then the full pattern of firing in this area will encode a set of parameters for eyes. The value of these parameters will determine some of the synaptic input on an assemblage of deep pyramidal cells, and thus it will modulate part of the signal being sent on the descending pathway. Therefore, instead of having one signal on this pathway that says 'eye' and produces a fixed template response in the lower area, there will be a family of varying signals, with spikes of varying rates and phases, representing eyes with different values of

the eye parameters. Each such signal will stimulate the lower area differently, producing the same effect as a flexible template with built-in parameters. Further, it is logical to suppose that the strengths of the synapses of other neurons on the dendrites of the deep pyramidals in the higher area are the place where the limits of this allowable variation is stored. These limits may be learned by gradual modification of these synaptic strengths, presumably by the presentation of multiple examples and by some mechanism which stores not just their mean but their variance in some form.

An interesting proposal for a specific way of representing and learning the variances of natural categories, as well as predicting the parameters for best fits, is being developed by Poggio and collaborators (Poggio and Girosi 1990; Poggio 1990). They hypothesize that both probability distributions for membership in a category and the values of associated parameters, as functions of a vector of features, may be approximated by a family of functions, called 'Hyper basis functions', of the form:

$$f(\mathbf{x}) = \sum_{\alpha=1}^{N} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|_W^2) .$$

Here $\mathbf{x}$ is the vector of features, $\mathbf{t}_\alpha$ are the exemplars from which the function has been learned, G is a function like a multi-dimensional Gaussian, the subscript W on the norm is a weighting of the individual features (e.g. an inverse of a covariance matrix) and c_α are learned weights. They propose neural mechanisms for implementing the calculation of such f's as well as learning the weights. Developing algorithms of this kind for learning variances or some other measure of natural variability of exemplars and for clustering similar exemplars seems to me to be a central problem for neural net architectures.

Fodor and Pylyshin (1988) have raised the question of how neural nets can express composite concepts and can rapidly build new composite concepts which have never been entertained before. Our notion of flexible templates seems to incorporate a limited form of compositionality in a natural way. When a template is active, the values of its parameters naturally associate to that concept a set of qualifying properties, much like a noun phrase may be formed by a principle noun and a set of adjectives and clauses. When several templates are active, their parameters don't get confused: the two scenes "Black dog and white cat" and "White dog and black cat" correspond to two different states of mental activity. Such linking does not allow us to arbitrarily combine two different concepts, but only to form combinations when one already occurs as a dimension of variability of the other. In a more linguistic context, there could be a template for the action "hit", whose parameters included a description of the object hitting and for the object being hit.

6 Residuals

Flexible templates were a major improvement on templates but flexible templates also have problems. How

Expected symbol with smudge | Best match by template | Unexpected symbol

Fig. 3. Problems in recognition by template

does one judge whether or not to accept the fit of the template and decide that the signal does contain a valid instance of the class of objects in question? An early idea was to decide that the letter 'B' was present on a page when part of the writing was more like a 'B' than any other letter. This procedure can go wrong in two ways. The page might have an unusual character here, not matching any English character, say the 'there exists' sign ∃ of mathematics, and you shouldn't have accepted 'B' just because the symbol ∃ was closer to a 'B' than any other English character. Or the paper might have an 'R' partially obscured by a smudge making the whole shape a bit closer to an 'B' than an 'R' say (see Fig. 3). The point is that identification is only complete when you have analyzed *all ways in which the signal differs from the template* (after putting in optimal values for its parameters). These differences are what I call the *residuals*. Many things may be happening.

1. The residual may be so large that the template is plain wrong (e.g. you were trying to fit the eye template to a mouth), and this particular identification should be rejected.
2. It may be that a definite part of the template is missing in the signal (e.g. an object in a scene is partially occluded by something in front of it), so you should accept the identification provided that the missing parts can be explained.
3. It may be that the signal contains something extraneous in addition to the template (e.g. while the sentence "Everything's fine at home" is being uttered, a child's scream breaks in), and again the identification should be accepted provided that the extraneous part can be explained.
4. Even when the correlation between the signal and the template is overwhelmingly large, so the identification is clearly correct, there are many situations when the residual contains very useful information about the world (see example in Sect. 8 below).
5. Finally, the residual may be 'noise' or 'clutter': unidentifiable, seemingly random stuff and one should then stop with the identification and not burden the rest of the algorithm further with its analysis or storage.

The moral here is that an animal should not rest until it has 'explained' the full set of signals coming to it from the world, as far as its past experience allows, and must also be able to recognize when the signal indicates – because of variations beyond the normal limits – something never encountered before. Ideas in this direction have been put forward by many people in

different contexts, for example, in computer vision by Pavlidis (1988).

The idea of residuals is closely related to concepts in the theory of robust statistics (Huber 1981). In robust statistics, one considers the problem of estimating the mean of a distribution from a sample, in the case where either the distribution itself has large tails or the sample is somehow corrupted. In both cases, the sample is likely to contain a few large outliers, which will cause large changes in the sample mean. Huber's solution is to explicitly identify the outliers and use the thinned sample to estimate more robustly the distribution's mean. Thus if 60% of a model or template fits a signal very well, one should explicitly mark the remaining 40% as outliers, and measure the goodness of fit of the remaining 60%. If this fit is good enough, it is usually stronger evidence than 90% of the template fitting the signal crudely.

7 Ascending pathways carry residuals

The last part of the proposal deals with the role of the superficial pyramidal cells. As before, we will only consider sensory areas in this section. Let us assume that a lower area B is interconnected with a higher area A. We always get two sets of connections and sometimes a third (with the single arrow):

Higher Area A *Lower Area B*
deep pyramidal cells $\Rightarrow$ *synapses in layers I and VI*
synapses in layer IV $\Leftarrow$ *superficial pyramidal cells*
superficial pyramidal cells $\rightarrow$ *synapses in layers I and IV*

Our proposal is that the loop with double arrows embodies an iterative algorithm that attempts to identify a specific higher level object in the lower level data. More specifically, the deep pyramidals of A send a signal to B containing the template for each predicted object O. Area B compares these templates to its blackboard, which gives its own present reconstruction of the world from its vantage point and computes a residual, a description of that part of the world which isn't expected or predicted. Its superficial pyramidals then send back to A this residual, a description of what doesn't fit A's prediction. The weights on its synapses in the higher area translate this residual into the higher level language. Then area A modifies the parameters in the flexible template to try to improve the fit and sends this back to B, and it may also hypothesize the presence of further objects in B's world. After a few turns, either a good fit is found and the residual is acceptably small or the hypothesis is rejected and area A turns to other previously suppressed hypotheses. In the ultimate stable state, the deep pyramidals would send a signal that prefectly predicts what each lower area is sensing, up to expected levels of noise, and the superficial pyramidals *wouldn't fire at all*[5]. At the other extreme, if you wake up in a strange place with no expectations or are totally surprised by something, then the algorithm starts with a clean slate in area A. Then B just sends its whole picture of the world to A which excites some possible higher level objects. At each stage, A writes on its blackboard its best guess in its language (objects and their parameters) about the identity of the higher level objects found in B's picture.

How do the extra descending paths from the higher area A to lower area B fit into our theory? Various even higher areas C_i are all predicting what A sees and this explains all but some residual of A's picture. A can tell the higher areas C_i about these unexplained features, to try to find a top-down explanation of them, *and* it can tell lower areas such as B about them. If the superficial pyramidals express the residual part of the world picture of higher area A, then the extra descending paths would carry such a message to B. Their effect could be to modify the world picture of area B, weakening the evidence on which these conclusions of A were based, pushing the lower area B to seek alternate parses of its data and to explain away the residual on a bottom-up basis. Note that this is different from sharing with B the reconstruction of the world which area A is entering: such sharing of conclusions can be accomplished through the thalamus, on which these conclusions are written.

My description is only the beginning of an algorithm for processing sensory input like a visual signal. But I have convinced myself of its plausibility by analyzing particular complex scenes of the world and seeking a 'rational reconstruction' of the process that the brain might follow in finding the correct semantic high-level interpretation. Such an approach has been followed by Cavanagh (1991), who analyzed recognition of faces in extreme lighting conditions producing dark shadows and confusing contours. His conclusion is that an algorithm very similar to our template/residual loop is the most likely possibility. What is most striking to someone who has experimented with small algorithms in computer vision – which operate without human prompting – is that any system of this type working on real visual input could be stable, could reliably integrate multiple small clues and find the *one* combination of hypotheses which explains the whole image. Nonetheless, I am proposing that a large number of independent loops, each looking for its pet structure in a lower level blackboard, working simultaneously on low and high levels, can in real world situations converge rapidly to the correct solution, without huge oscillations and without creating fanciful high level images unconnected to reality.

8 Comparison with other top-down/bottom-up theories

The proposed sketched above has many similarities both to the 'adaptive resonance theory' of Carpenter and Grossberg (1987), the 'counter-current processing model' of Deacon (1988), Poggio's Hyper basis function theory (1990) and Roll's theory of backprojections in cortex (1990).

[5] In some sense, this is the state that the cortex is striving to achieve: perfect prediction of the world, like the oriental Nirvana, as Tai-Sing Lee suggested to me, when nothing surprises you and new stimuli cause the merest ripple in your consciousness

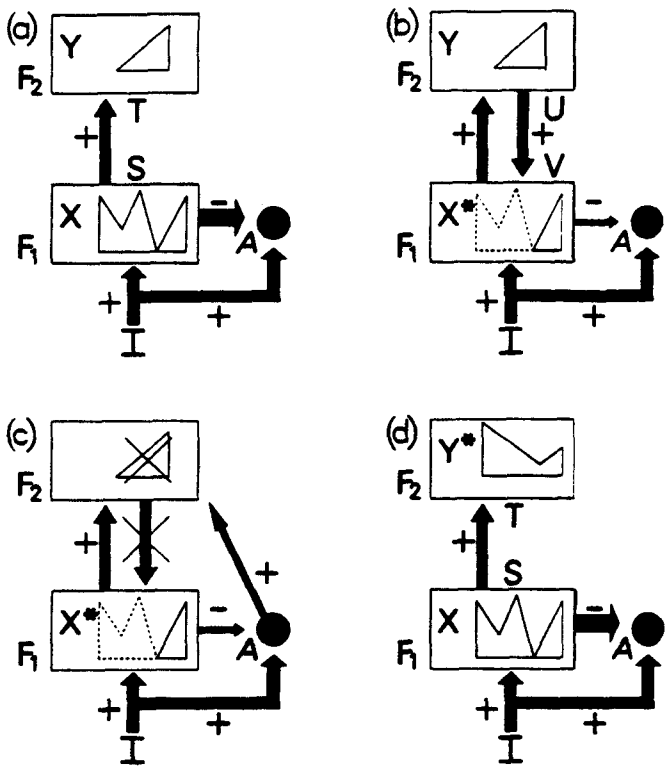

Fig. 4a–d. The "Adaptive Resonance Theory" of Carpenter–Grossberg (Carpenter and Grossberg 1987)

Thus Carpenter and Grossberg's theory is summarized in Fig. 4, and it works like this: F_1 and F_2 are two cortical areas projecting to each other. A pattern of activity X in F_1 (shown by the symbolic pattern in (a)) evokes a signal S on the bottom-up pathway to F_2. S stimulates a pattern of activity T among all the stored categories into which X might be classified, and, by a winner-take-all algorithm, the best fitting category Y is selected. The pattern Y in F_2 evokes a signal U on the top-down pathway, which stimulates a pattern of activity V in F_1, the template or 'learned expectation'. V and X combine to form X^*: either X^* is close to X, in which case the network stabilizes and classifies X as being an instance of Y, or else mismatch occurs. Panel (b) in the figure shows the latter, and, in this case, it results in an 'arousal burst' from module A which inhibits Y in a long-lasting way (panel (c)). Now the same pattern of activity T on F_2 no longer selects Y but the second best matching category Y^*, which is in turn compared with X, etc.

We see that ART posits a recursive calculation in a top-down/bottom-up loop which is very similar to ours. The first major difference, however, is that in ART, the templates store some kind of mean or median representative of each learned category, and make no attempt to explicitly encode the variation within a category as in the work of Poggio et al. (1990). In Poggio's analysis, a suitable set of exemplars for each category are stored, and used to generate a smooth function which approximates the probability that a new feature vector input should be interpreted as a member of the same category. As explained above, I feel that this is essential to any successful recognition algorithm. Moreover, along with storing variances, the degree of mismatch should not be merely a number, whose size determines whether or not to seek a new category, but a signal representing what does not match. This is the second major difference, and leads to our idea of residuals. Making explicit such residuals allows the higher area to seek complex explanations of the input in which several templates are superimposed.

The following vastly simplified illustration may explain why I feel storing variances and describing residuals is essential in real-life situations. Suppose two numerical features x and y are computed from an olfactory stimulus, and suppose the world contains two animals A and B. Suppose that the smell of A excites x and y roughly equally, but that the smell of B excites x but not y. In a noisy world, we should never say the B's smell doesn't ever excite y, but rather something like: in the presence of B alone, the value of y is almost always at most 1/20th that of x. Finally, suppose A is dangerous while B is not. Then suppose the input has values $x = 5$, $y = 1$. The template for A is $x = y =$ any positive value (the smell may be strong or weak depending on the proximity of A), and the template for B is $x =$ any position value, $y = 0$. Clearly, we get a much better correlation of the input with the template for B, with a suitable parameter put in. But, knowing the variation expected in the smell of B, we see that there is non-trivial residual. The best fit by B alone might be $x =$ about 5, $y = 0.25$, with a residual $x =$ unknown, $y = 0.75$. The residual can be fitted with the smell of A, and we recognize the presence of danger (see Fig. 5). Note that to carry out this procedure, we need both to explicitly encode the variability of B's smell and to separate the relatively small unexplained part of the input from the dominant part explained by the first template. Situations of this type occur more often than not in the analysis of real visual data for instance.

Deacon has also proposed a theory of cortical processing based on top-down/bottom-up loops, that he calls 'counter-current' processing. Like ours, his theory is motivated by the laminar assymetry between ascending and descending cortico-cortical pathways. He proposes that a form of relaxation between the information in two mutally connected areas takes place, the higher areas sending down foci of attention, expectation and associated imagery, while lower areas send up perceptual details and recognized patterns. He does not assign as precise a computational role, however, to the two streams as I do, but develops an interesting metaphor of two fluids moving through adjacent tubes in opposite directions, where some quantity like heat diffuses from one to another at all points of contact of the tubes. Finally, Rolls has discussed the bottom-up/top-down loops in cortex in connection with memory and the learning of categories and has a primitive neural net simulation of this loop. He stresses the importance of separating two stimuli which are close in one sensory modality, but which are learned to be very different from experience in other modalities. Both

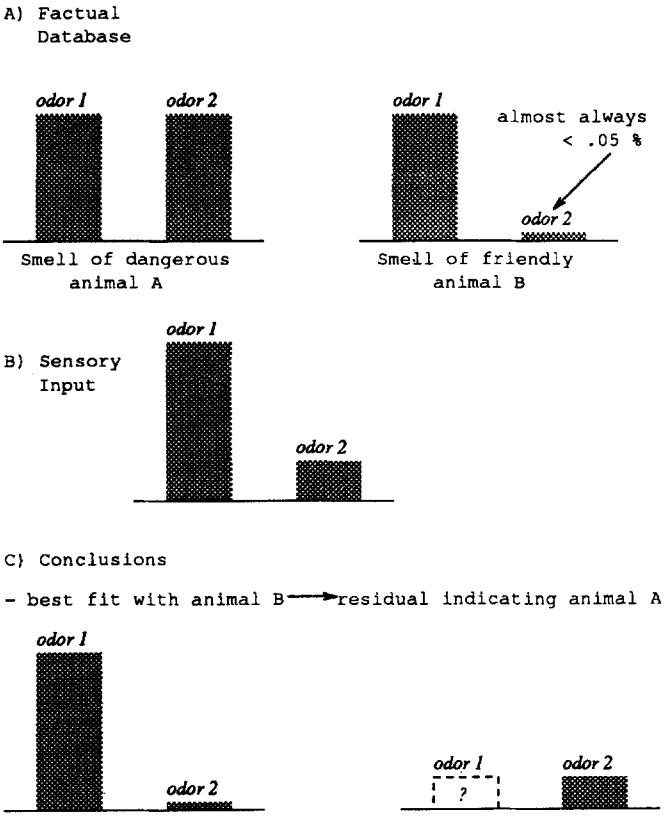

Fig. 5A–C. The importance of variances and residuals

Deacon and Rolls analyze the entorhinal/hippocampal complex at the top end of the cortical area hierarchy and its role in the formation of memories, which I am excluding from this paper.

9 Managing the top-down/bottom-up loop

If the top-down phase and bottom-up phase of a pattern recognition algorithm are to work effectively together, it would seem necessary to coordinate them. In sequential, von Neumann architecture terms, one could think of the loop consisting of *a*) one 'cycle' of computation in the lower area, *b*) passing the data up to the higher area, *c*) one 'cycle' there and finally *d*) passing the data back down. Then the brain would operate by a relaxation algorithm, in which the loop is repeated until it stabilizes. The brain being, by nature, highly parallel, and there being not just one pair of lower/higher areas but many, it is more reasonable to imagine the lower and higher area working at once, and then exchanging their data.

Very recent experiments by Gray and Singer (1989) have discovered that strong local oscillations with a 20–30 ms period (corresponding frequency 35–50 Hz) accompany periods of intensive computation in at least some areas of cat cortex (V1 and V2). The oscillation may be detected in the mean local electrical field, or in single cell recordings from a large number of individual cells. Freeman (cf. the review Freeman and Skarda 1985) has found similar, somewhat faster, oscillations in rabbit olfactory bulb caused by alternating bursts in pyramidal cells and interneurons, but the bulb, like the cerebellum, differs from cortex in having an order of magnitude more interneurons than output neurons, suggesting quite different computational principles.

It seems logical to propose that these oscillations are caused by or synchronize calculations in the ascending/descending pathway loop. (The same suggestion has been entertained by Hubel and Livingstone – oral communication.) It may be that iteration in this loop would be unstable if the top-down and bottom-up phases occurred asynchronously. This would predict that oscillations like those found in V1 by Singer will be found in every area of the cortex and that successive waves of top-down signals and bottom-up signals occur at specific phases of the local oscillation. Bursts of this oscillation will coincide with active local computations using this loop. This makes a very specific prediction: that if simultaneous recordings are made from deep pyramidals in a higher sensory area and superficial pyramidals in a lower sensory area which project to each others 'columns', i.e. to near each other, then bursts in the two populations will be phase locked with each other, and with the local mean electrical field. Allowance must be made for the fact that signals are by no means simultaneous firings of all pyramidals of each class: the information is precisely in which ones are firing and probably in timing differences of their individual spikes too. But suitably averaged (as in the recordings that demonstrate the oscillation to begin with), I would expect to see this synchronization of remote neurons.

There are other parts of our theory in which synchronization is needed. The time buffering in auditory and motor cortex presumably needs some kind of pace maker. But this would be much slower, e.g. 1 to 10 Hz. The time scale of our top-down, bottom-up loop must be much faster or the brain would never get anything done. The experimental finding of 35–50 Hz seems about right: in half the period, 10–15 ms, the local intracolumnar circuits of the cortex should have time to do non-trivial calculations, and there is time for half a dozen iterations of the loop before the books close on interpreting a stimulus.

An intriguing possibility is that the claustrum plays some role in modulating the 'top-down', 'bottom-up' calculation between various cortical areas. The claustrum is a relatively small subcortical nucleus that is located like a seventh layer of the cortex just beneath a certain cortical area, the insula, but separated from it by a thin layer of white matter, the extreme capsule. It is connected to almost the whole neocortex, but *not topographically*! This is a major exception to the pattern for other connections and means that if two cortical areas A and B projects to parts A' and B' in the claustrum, than A' and B' often overlap. In fact, Pearson et al. (1982) have made the following generalization on the basis of extensive primate studies:

- A' and B' overlap if and only if the cortical areas A and B are directly connected by cortico-cortical pathways.

It should also be noted that the claustrum is an evolutionarily conservation form, being present and similarly connected in the most primitive mammals. These facts makes it look likely that the claustrum is connected to the operation of these reciprocal pathways in some essential way. Now the claustrum seems to have too few neurons to play a role in the substance of the calculation taking place in the loop, but it is ideally situated to modulate the relaxation algorithm between the areas in some way, e.g. initiating and terminating it or in some way maintaining its stability (see also Crick and Koch (1990) for a related proposal).

10 Mental images

Another enticing speculation is to consider the action of the brain in a purely introspective state. In the course of reflecting about some problem, we can block out the actual stimulus being received by our senses, or we can close our eyes. At that point, all the neural machinery for sensory processing is available for thought. I want to conjecture that the process of thinking things through often involves writing in a purely top-down mode on the active blackboards of low level areas, and using the various reciprocal pathways to better understand a situation or problem which is not physically in front of us. This can be done by the deep pyramidals cells which will evoke a template in the activity of the lower area, and thus write this template on its blackboard. I want to propose that this is exactly what we do when we form a mental image of some object. This system of thinking can be applied, e.g. to work out tricky things about the three-dimensional geometry (can we carry a piano up the apartment stairs), or to work out more abstract problems using amorphous objects as tokens for parts of some situation.

The mental rotation experiments of Shepard and collaborators (Shepard and Cooper 1982), suggest that analog, continuous, real-time rotation is often performed on mental images. A natural interpretation of their results in the present context is that this step-by-step transformation is carried out by the top-down, bottom-up loop between cortical areas. In many ways it is analogous to the relaxation algorithms using the same loop by which parameters in a template are iteratively adjusted to achieve a better fit with a stimulus. In this case, however, a mental image which is a rotated version of one stimulus is iteratively adjusted to achieve a better fit with another stimulus.

Moreover, it is also known that the brain is working intensely during dreams without any sensory input and that the thalamus is quite active. The visual images present during dreams would seem to be stimuli evoked purely by top-down pathways. During dreaming sleep, the brain also receives diffuse cholinergic stimulation from brain stem nuclei via so-called 'PGO waves' (Mamelak and Hobson 1988). But the vivid, often realistic though bizarre, images of dreams would seem to require something like our template generating deep pyramidal neurons. I hypothesize that the same mechanism that gives rise to mental images when awake drives the formation of dream images. Their bizarreness may result from the evocation of multiple top-down images simultaneously for some as yet unknown cognitive/emotional function.

11 Possible tests

A much debated property of V1 neurons is that of 'end-stopping'. Neurons with this property fire in the presence of bars or edges, moving or still, with a fixed orientation and a fixed location provided they are not too long. That is, the stimulus must be contained in a certain receptive field and it mustn't continue outside this field. Zucker and collaborators (Dobbins et al. 1987) have hypothesized that this is due to the neuron computing the curvature of the bar or edge, and that it would fire strongly if the bar or edge continued but turned with approximately a specific curvature. A radically different hypothesis is that the neuron does fire briefly to a longer line, but that as soon as top-down signals from V2 incorporate this long line into a global segmentation of the scene, the line is accounted for and the firing stops. In other words, its firing indicates that the line is unexpected, and not part of a coherent global pattern. A short line never fits into such a pattern and firing continues: it remains a residual. This theory would predict that superficial end-stopped cells, i.e. those in layer II and III, would be responding to residuals. This predicts that their end-stopping would not be absolute: they would have a transitory response to longer edges, which would be inhibited as soon as the V1 $\Rightarrow$ V2 $\Rightarrow$ V1 loop kicks in (say 20 ms). It would further predict that deep end-stopped cells respond more consistently as in Zucker's theory to some property of the stimulus.

More generally, a plausible prediction of the theory is that many of the responses of superficial pyramidal cells should be transitory. The idea is that when everything being sensed is predicted or explained by the high levels' models of the state of the world, there are no more residuals to send upstream. In a calm state of meditation, for instance, their overall activity would diminish substantially.

Another conjecture would be similar to the classic experiment of DeValois et al. (1979) in which the retinotopy of V1 was revealed by fixing a visual stimulus on the retina, injecting the animal with radioactive glucose taken up in metabolism, killing the animal after a short period and examining the pattern of radioactivity present post mortem in V1. I would propose stimulating strongly a deep pyramidal cell in an area like V2 or V4, connected to V1, while again marking cell activity via a radioactive metabolite. If V4 is concerned with shape recognition, template-like shapes should appear in V1. Precisely because V4 is not finely retinotopic, the pattern of activity in V1 would be extended and not precisely localized.

References

Brodal A (1981) Neurological anatomy. Oxford University Press, Oxford

Carpenter G, Grossberg S (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. Comp Vision Graphics Image Proc 37:54–115

Cavanagh P (1991) What's up in top-down processing. In: Gorei A (ed) Representations of vision. Camb. University Press, Cambridge, pp 295–304

Crick F, Knock C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci (in press)

Deacon T (1988) Holism and associationism in neurophyschology: an anatomical synthesis. In: Perecman E (ed) Integrating theory and practice in clinical neuropsychology. Erlbaum, Hillsdale NJ

Deacon T (in preparation) Laminar organization of frontal cortico-cortical connections in the monkey brain

DeFelipe J, Jones E (1988) Cajal on the cerebral cortex. Oxford University Press, Oxford

Desimone R (1991) Face selective cells in the temporal cortex of monkeys. J Cogn Neurosci 3:1–8

DeValois KK, DeValois R, Yund EW (1979) Responses of striate cortex cells to grating and checkerboard patterns. J Physiol (London) 291:483–505

Dobbins A, Zucker S, Cynader M (1987) Endstopping in the visual cortex: a neural substrate for calculating curvature. Nature 329:96–103

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in primate cerebral cortex. Cerebral Cortex 1:1–47

Fischler M, Elschlager RA (1973) The representation and matching of pictorial structures. IEEE Trans Comp 22:67–92

Fodor J, Pylyshin X (1988) Connectionism and cognitive architecture. Cognition 28:3–71

Freeman W, Skarda CA (1985) Spatial EEG patterns, non-linear dynamics and perception. Brain Res Rev 10:147–175

Gray C, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci 86:1698–1702

Huber P (1981) Robust statistics. Wiley, New York

Ito M (1984) The cerebellum and neural control. Raven Press, New York

Mamelak A, Hobson JA (1988) Dream bizarreness as the cognitive correlate of altered neuronal behavior in REM sleep. J Cogn Neurosci 1:201–222

Pavlidis T (1988) Image analysis. Ann Rev Comput Sci 3:121–146

Pearson RCA, Brodal P, Gatter KC, Powell TPS (1982) The organization of the connections between the cortex and the claustrum in the monkey. Brain Res 234:435–441

Perkel DJ, Bullier J, Kennedy H (1986) Topography of the afferent connectivity of area 17 in the Macaque monkey. J Comp Neurol 253:374–402

Poggio T (1990) A theory of how the brain might work. In: The Brain. Proc Cold Spring Harbor Symp 55

Poggio T, Girosi F (1990) A theory of networks for learning. Science 247:978–982

Primrose D, Strick P (1985) The organization of interconnections between the premotor areas of the primate frontal lobe and the arm area of the primary motor cortex. Soc Neurosci (abstr) 11:1274

Rock I (1984) Perception. Sci. Am. Books, New York

Rockland KR, Virga A (1989) Terminal arbors of individual 'feedback' axons projecting from area V2 to V1 in the macaque monkey. J Comp Neurol 285:54–72

Rolls ET (1990) The representation of information in the temporal lobe visual cortical arteas of macaques. In: Eckmiller R (ed) Advanced neural computers, Elsevier, New York Amsterdam pp 69–78

Shepard R, Cooper LA (1982) Mental images and their transformations. MIT Press, Lancaster

Van Essen DC, Newsome WT, Maunsell JHR, Bixby JL (1986) The projections from striate cortex to areas V2 and V3 in the Macaque monkey. J Comp Neurol 244:451–480

Winfield DA, Gatter KC, Powell TPS (1990) An electron microscopic study of the types and proportions of neurons in the cortex of the motor and visual areas of the cat and rat. Brain 103:245–258

Yuille A (1991) Deformable templates for face recognition. J Cogn Neurosci 3:59–70

Dr. D. Mumford
Mathematics Department
Harvard University
1 Oxford Street
Cambridge, MA 02138
USA