The recent excitement about neural networks

Francis Crick

The remarkable properties of some recent computer algorithms for neural networks seemed to promise a fresh approach to understanding the computational properties of the brain. Unfortunately most of these neural nets are unrealistic in important respects.

THERE has been a lot of excitement recently about neural nets. A new algorithm has produced quite simple nets that perform surprisingly well. A thick two-volumed work, *Parallel Distributed Processing*[1], has been a best-seller, read enthusiastically by psychologists, computer designers and physicists. Even undergraduates are now designing new networks. The interested spectator may well wonder what it's all about. What are neural nets? How do they work? And what, if anything, do they tell us about the brain?

Neural nets are composed of 'units' that have some of the properties of real neurons[2]. That is, each unit has many inputs, some of which excite it and some of which inhibit it. The unit usually takes a weighted sum of all the inputs and puts out a single output, down the 'axon', if the weighted sum exceeds some threshold. In very simple nets the output may have only two states (0 for inactive, 1 for active). Other nets may use units with graded output, in which the output scalar can be loosely thought of as representing the average spike rate of a neuron. Such parallel nets are not, in most cases, built using analogue components. Instead they are usually simulated, rather laboriously, on a computer.

Everyone would agree that to understand the brain we must know how groups of neurons interact. As we are sure that many synapses are plastic, that is, they can change their strength with experience, it is also important to know the abstract rules imposed on such changes by the intricate biochemistry of a neuron. Thus a net is characterized by the properties of the units that make it up, the way they are connected together, and the algorithms used to change the strength of those connections.

Memory

What have we learnt from such theoretical models? An early discovery was that memory could be stored in a way very different from memory storage in a standard digital computer (see ref. 3 for a recent historical review). This is perhaps not too surprising. A typical computer is made with very fast components, each having rather few inputs and capable of sending a pulse-coded message. Part of each message is the 'address' that indicates where a particular memory can be stored, and part is the information to be stored. The operation of the computer is largely serial.

The brain is different in almost every respect. Neurons are slow, operating in the millisecond time range, and typically have many hundreds or thousands of inputs. Although many of them produce action potentials or 'spikes' whose distribution in time is not completely random, there is no obvious sign of precise pulse-coded messages. Moreover the parts of the brain seem to be highly parallel in their operation[4]. How then can an assembly of neurons—a net—store memory?

Notice that there are three aspects of processing a memory: putting it into the net, storing it there over time and retrieving it when required. Leaving aside immediate or working memory, which may be rather different, it is widely believed that the first and third operations (the in and the out) require neural activity, whereas the second one—the long-term storage—does not. The memory, so the gospel goes, is embedded in the strength of the numerous connections or synapses in the network. Most neural nets also have this character.

What has to be stored in a net is the capacity to produce a particular pattern of activity in a group of units. By suitably adjusting the strengths of all the synapses, using a simple, local rule, a net can produce a pattern, given a suitable 'clue'. The clue can be any smallish part of the desired pattern. This can be done especially easily if the net feeds back on itself. Given a part of the input pattern, the net, by self-excitation, will regenerate the whole pattern. Such a system is called content-addressible, because any part of the pattern can act on the clue, which provides the address. Moreover, nets of a reasonable size can store several patterns. If they are sufficiently distinct the patterns will not interfere with each other. Thus the system is distributed as one memory is distributed over many synapses; superimposed, because one synapse can be involved in several memories; and robust, because altering a few synapses degrades the performance very little[5].

Such nets are usually simple, having a single layer of units. Moreover they are usually unsupervised. That is, there is no teacher to tell the net how to adjust its output to make it resemble the desired one. The net learns by using an algorithm based on an idea of Hebb[6]. This is a local algorithm as it depends only on the activity near that particular synapse: roughly, the synapse is strengthened if it receives an input signal on the presynaptic side, together with some indication of activity, such as the unit firing, on the postsynaptic side.

Back propagation

Memory, however, is not the only thing the brain needs to achieve. Another type of useful net is one that can extract categories. That is, it must search for regularities and correlations in the incoming signals and try to embody them in some way in its performance. It turns out that a single layer of units without feedback has severe limitations[3]. Even if each unit is told after each trial whether it should have fired faster or slower, a procedure known as supervised learning, it cannot be trained to perform even quite simple operations. The classic example is the exclusive OR (A, or B, but not both A and B). This can easily be done if a net of several layers is allowed. Unfortunately this leads to a problem: of all the various synapses, which ones should be adjusted to improve performance? This is especially acute if the synapses lie on several different layers of neurons.

The recent excitement has sprung mainly from a neat algorithm which solves this problem surprisingly well[7]. The full name of the algorithm is 'the back propagation of errors' but it is often called back-prop for short. It can be applied to any number of layers, although only three layers are usually used: an input layer, a middle layer (referred to as the hidden units) and an output layer. A unit in each of the first two layers connects to all units in the layer immediately above (Fig. 1). There are no reverse connections or sideways connections—a simple net indeed. Each unit forms the usual weighted sum of its inputs and emits a graded output.

The net works as follows. It is first set up with random connections. A particular input is then given to it, which produces activity in the output layer. A teacher, who knows what the response of each output unit should be for that particular input, indicates to each unit the size and sign of its error. For a theoretical model, the teacher is usually the person designing the net. In the brain the teacher is presumed to be another part of the brain. The error signals are used to adjust the weights of all the connections to the top or output layer and this information is 'back-propagated' to the hidden units in the middle layer. They use this information to adjust their synapses, which come from the axons of the input units. The exact way this is done conforms to common sense, but non-mathematicians often find the algebra rather daunting, so I will not attempt to describe it here[7].

It can be shown that, in effect, these adjustments are equivalent to a method of gradient descent[7]. This means that the algorithm used makes a small adjustment to the strength of each synapse, in such a way that each alteration reduces the total error in the performance of the net. Applied repeatedly, this necessarily leads to a minimum in the total error. It is well known that such methods run the risk of being trapped in a local minimum[7], which may be far above the global minimum. But in these nets this rarely happens, probably because of the many units involved and because of their graded response. Moreover the absolute global minimum is usually not required, only a sufficiently deep one.

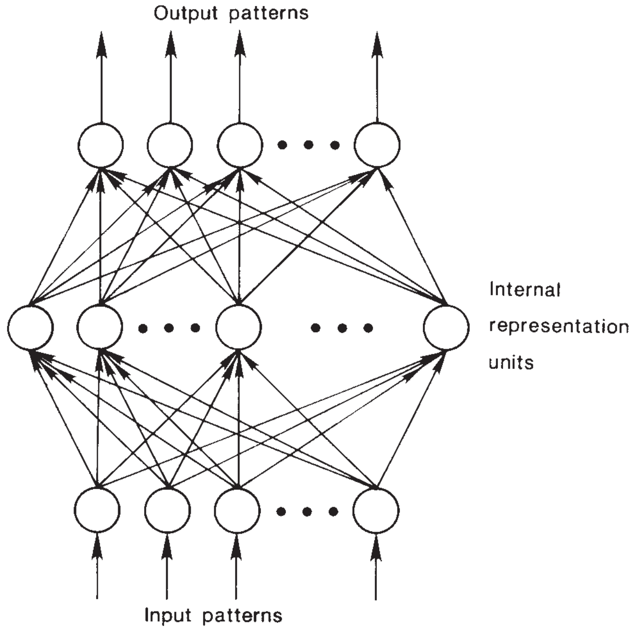

Fig. 1 A multilayer network. The internal representation units are sometimes termed hidden units. The information coming to the input units is recoded into an internal representation and the outputs are generated by the internal representation rather than by the original pattern. Input patterns can always be encoded, if there are enough hidden units, in a form such that the appropriate output pattern can be generated from any input pattern. (Reproduced, with permission, from ref. 1).

NETtalk

The results that can be achieved with such simple nets are astonishing. A striking example, due to Sejnowski and Rosenberg[8], is called NETtalk, whose task is to learn to pronounce English. It takes an English text as its input (Fig. 2) and its output is fed to a machine that can produce voice sounds. At first, having only random connections, it babbles, but gradually, as training proceeds, it starts to speak more intelligibly. Eventually, when tested with a text it has never seen before, it produces quite passable English speech, with about 90% accuracy (it could never be 100% correct because pronunciation depends somewhat on context in English, and the network knows nothing of meaning). Thus it has learned the rules of English pronunciation, which are notoriously not straightforward, in a tacit manner, from examples only, and not because the rules have been explicitly embodied in some program.

How does it do this? It is relatively easy, once such a net has been trained, to examine the 'receptive fields' of all the units in the hidden layer (by 'receptive' field is meant those features of the environment to which the unit responds). The results are remarkable[8]. The required information about categories is distributed across these neurons, because all information has to pass through the hidden layer, but not by any means in a random fashion. The hidden units have latched on to significant aspects of English speech, such as the difference between vowels and consonants, and indeed the different subclasses of these categories.

Further studies[8] have shown that it is important to have the correct number of hidden units. With too few, the net simply cannot do the job. Too many, and, although the net works a little better, it does not generalize as well—it fails on English text it has not been trained on. It has become, as it were, merely a look-up table. But given the correct number of hidden units it will extract significant categories, which it can use successfully on untested material of the same general kind.

This is not the only example. Other striking applications are to the problem of deducing the secondary structure of a protein from its amino-acid sequence[9], the distinction between rocks and unmentionable metal objects on the sea bottom[10] and the derivation of shape from shading[11].

The last example also teaches us a further lesson. The units produced in the shape-from-shading problem turn out to be rather like edge or line detectors in the visual cortex. This alerts us to the fact that the receptive field of a neuron, by itself, does not necessarily tell us what its main function is. This also will depend on where such a neuron projects, its 'projective field' as it has been called[11].

Neural nets and the brain

It is hardly surprising that such achievements have produced a heady sense of euphoria. But is this what the brain actually does? Alas, the back-drop nets are unrealistic in almost every respect, as indeed some of their inventors have admitted. They usually violate the rule that the outputs of a single neuron, at least in the neocortex, are either excitatory synapses or inhibitory ones, but not both[12]. It is also extremely difficult to see how neurons would implement the back-prop algorithm. Taken at its face value this seems to require the rapid transmission of information backwards along the axon, that is, antidromically from each of its synapses. It seems highly unlikely that this actually happens in the brain. Attempts to make more realistic nets to do this[13], though ingenious, seem to me to be very forced. Moreover the theorists working on

1. McClelland, J. L., Rumelhart, D. E. & PDP Research Group (eds) *Parallel Distributed Processing* (MIT, Cambridge, Massachusetts, 1986).
2. Rumelhart, D. E., Hinton, G. E. & McClelland, J. L. in *Parallel Distributed Processing* (eds McClelland, J. L., Rumelhart, D. E. & the PDP Research Group) 45–76 (MIT, Cambridge, Massachusetts, 1986).
3. Cowan, J. D. & Sharp, D. H. *Proc. Am. Acad. Arts Sci.* **117**, 85–121 (1988).
4. Van Essen, D. C. in *Cerebral Cortex* Vol. 3 (eds Peters, A. & Jones, F. G.) 259–324 (Plenum, New York, 1985).
5. Hopfield, J. J. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
6. Hebb, D. O. *Organization of Behavior* (Wiley, New York, 1949).
7. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. in *Parallel*

the subject are so remote from actual neurons that they have been cavalier in omitting one type of unit altogether. Obviously there should be a unit to compare the output of each output neuron with the signals from the teacher, in order to calculate the error (in back-prop nets this is done by the computer). Such a set of neurons, if they exist, should have novel properties and would be worth looking for, but there is no sign of back-prop advocates clamouring at the doors of neuroscientists, begging them to search for such neurons.

There are many return pathways in the brain[4], but we do not yet know if any of them act as one of the proposed teachers. Notice that to send separate teaching signals to each output neuron, a pathway must carry a lot of detailed information. We do indeed see diffuse pathways, such as that from the locus coeruleus, but one such neuron sends much the same signal to many parts of the brain, so that the information it can convey is somewhat limited and certainly would not be enough to control back-propagation. It may of course be used to tell the system when something is worth remembering. Models along these lines have been suggested, for example by Barto[8].

Another problem is that though the back-prop algorithm can be generalized to a system with several successive hidden layers, it becomes extremely cumbersome. An ingenious way round this, suggested by Hinton[15], is to train the net to make its output exactly the same as its input. This still allows the hidden layer to extract categories. In such a system the true output is taken directly from this hidden layer and made the input for the next set of nets. This allows any number of nets to be stacked on top of each other. If this were combined with a diffuse signal to indicate that something worth remembering has occurred, it begins to have some faint resemblance to what we see in the brain, but we are still stuck with the problem of how to implement back propagation realistically. Obviously what is really required is a brain-like algorithm which produces results of the same general character as back propagation. Another objection is that the back-prop algorithm is too slow, though it is not easy to make this argument quantitative in a realistic way. In any case modifications are now being suggested which can work rather faster[16].

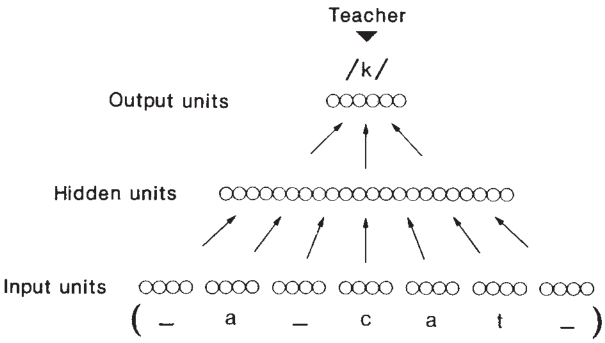

Fig. 2 A schematic drawing of the NETtalk network architecture, a specific example of the general scheme in Fig. 1. A window of seven 'letters' in an English text ('a cat' here) is fed to an array of 203 input units. Information from these units is transformed by an intermediate layer of 80 'hidden' units to produce patterns of activity in 26 output units. The connections in the network are specified by a total of 18,629 weight parameters, including a variable threshold for each unit. Reproduced, with permission, from ref. 8.

Most of these neural 'models' are not therefore really models at all, because they do not correspond sufficiently closely to the real thing. I have suggested elsewhere[17] that they be called 'demonstrations'. They refute the claim that it is impossible for any neural net to act in such-and-such a way, but they may not perform in just the way the brain does. In another context they might reasonably be referred to as existence proofs. As such they have a certain use. The back-propagation algorithm can be used to generate a set of useful synaptic weights. This may not be the way the brain arrives at them—it may use some other more realistic algorithm—but the resulting receptive fields may, in their general character, be somewhat similar to those the brain has arrived at. This can be tested experimentally.

A good example of this approach is the work of Zipser and Andersen[18], modelling a subset of posterior parietal neurons in the macaque monkey. Another promising case is the modelling of the vestibulo-ocular reflex[19]. In both, the character of the hidden units produced by back propagation is somewhat similar to what is found in electrophysiological recordings. Nevertheless, as far as the learning process is concerned, it is unlikely that the brain actually uses back propagation. In spite of this, back propagation has been greeted with widespread enthusiasm.

Why the excitement?

How has this curious situation arisen? Apart from a few enthusiasts, most theorists do not believe that, for example, children really learn to speak using a single, simple back-prop network inside their heads. Why, then, are such models considered not only useful, but exciting?

To understand this we have to look at the structure and history of the various disciplines involved. It comes as a surprise to neuroscientists to discover that many psychologists, linguists in particular, have very little or no interest in the actual brain, or at least what goes on inside it. The brain, they feel, is far too complicated to understand. Far better to produce simple models which can do the job in an intelligible manner. That such models may have little resemblance to the way the brain actually behaves is not seen as a serious criticism. If it describes, in a succinct way, some of the psychological data, what can be wrong with that? Notice, however, that by using such arguments, one could easily make a good case for alchemy or for the existence of phlogiston.

The position is complicated by the fact that there is another application for network models: to assist in the design of novel, highly parallel computers. For this it makes no difference how the brain works and it is in this general area that most advances are now being made. Eventually, of course, the computer circuitry will have to be embodied in some sort of chip, which will bring its own design problems. The back-prop algorithm can be used to develop a good set of weights for special purpose chips, though eventually more versatile chips will be needed, with modifiable connections. Meanwhile, so the argument goes, why not develop networks and algorithms to see which systems perform best. With luck this may give theorists some experience of how complex

Distributed Processing (eds McClelland, J. L., Rumelhart, D. E. & the PDP Research Group) 318–362 (MIT, Cambridge, Massachusetts, 1986).
8. Sejnowski, T. J. & Rosenberg, C. R. *Complex Systems* **1**, 145–168 (1987).
9. Qian, N. & Sejnowski, T. *J. molec. Biol.* **202**, 865–884 (1988).
10. Gorman, R. P. & Sejnowski, T. J. *Neural Networks* **1**, 75–89 (1988).
11. Lehky, S. R. & Sejnowski, T. J. *Nature* **333**, 452–454 (1988).
12. Crick, F. H. C. & Asanuma, C. in *Parallel Distributed Processing* (eds McClelland, J. L. Rumelhart, D. E. & the PDP Research Group) 333–371 (MIT, 1986).
13. Zipser, D. & Rumelhart, D. E. in *Computational Neuroscience* (ed. Schwattz, E.) (MIT, in the press).
14. Barto, A. G. *Human Neurobiol* **4**, 229–256 (1985).
15. Hinton, G. E. *Artificial Intelligence* (in the press).
16. Fahlman, S. E. Technical Report CMU-CS-88-162, Carnegie Mellon, Computer Science Dept Publication CMU-CS-88-162 (Pittsburgh, Pennsylvania, 1988).
17. Crick, F. *What Mad Pursuit* (Basic Books, New York. 1988).
18. Zipser, D. & Andersen, R. A. *Nature* **331**, 679–684 (1988).
19. Anastasio, T. J. & Robinson, D. A. *Neural Computation*, submitted.
20. Hopfield, J. J. & Tank, D. W. *Science* **233**, 625–633 (1986).
21. Churchland, P. S. & Sejnowski, T. J. *Science* **242**, 741–745 (1988).
22. Jacob, F. *Science* **196**, 1161–1166 (1977).
23. Ramachandran, V. S. in *Vision, Coding and Efficacy* (ed. Blakemore, C.) (Cambridge University Press, 1988).
24. *Trends in Neuroscience* **10**, 263–302 (1987).

non-linear nets behave in practice. Whether this experience can usefully be applied to models of the brain remains to be seen.

I also suspect that within most modellers a frustrated mathematician is trying to unfold his wings. It is not enough to make something that works. How much better if it can be shown to embody some powerful general principle for handling information, expressible in a deep mathematical form, if only to give an air of intellectual respectibility to an otherwise rather low-brow enterprise.

At the bottom, one has to realize that these various activities, though superficially similar, are of a radically different kind. Constructing a machine that works (such as a highly parallel computer) is an engineering problem. Engineering is often based on science but its aim is different. A successful piece of engineering is a machine which does something useful. Understanding the brain, on the other hand, is a scientific problem. The brain is given to us, the product of a long evolution. We do not want to know how it might work but how it actually does work. This has been called 'reverse engineering'—trying to unscramble what someone else has made—but as has been pointed out, it is reverse engineering[20] on the products of an alien technology[21].

And what a technology! Natural selection is not a clean designer. As François Jacob[22] has pointed out, evolution is a tinkerer. It has, broadly speaking, to build on what went before. It can take a simple, straightforward process, such as DNA replication, and embroider it with any amount of gadgetry to make it work a little better. It is opportunistic: anything will do as long as it works. Naturally it is constrained by both chemistry and physics, but this does not necessarily mean that its mechanism will embody deep general principles; the structure of the genetic code is a good example of this. It may prefer a series of slick tricks to achieve its aim[23]. Only a close inspection of the gadgetry will tell.

And this brings us to the crux of the matter. Why not look inside the brain, both to get new ideas and to test existing ones? The usual answer given by psychologists is that the details of the brain are so horrendously complicated that no good will come of cramming one's head with that sort of information. To which the obvious reply is, "If it's as complicated as that, how do you hope to unscramble its workings by a purely black-box approach, by merely looking at its inputs and outputs?"

By looking inside the brain we now strongly suspect that in important cases, at least in vertebrates' synaptic modification depends on the behavior of the NMDA-type glutamate receptor (see, for example, various articles in ref. 24). This receptor works on a slightly slower timescale than the other glutamate receptors. It will open only if it has received the neurotransmitter glutamate, or something like it, during the recent past, provided the local negative voltage of the cell membrane has become somewhat more positive than the normal resting potential as a result of other inputs. When it does open, it lets in a lot of calcium ions, thought to be one of the initial signals in the complicated process of synaptic modification. It is thus ideally suited to perform associative learning.

Models embodying behaviour of the NMDA receptor would be most welcome and indeed such work is already in progress (D. E. Rumelhart, personal communication). As the weight of an NMDA receptor depends on the activity of neighbouring synapses in altering the membrane potential, this opens the possibility of multiplicative interactions between synapses. We urgently need to know the exact location of NMDA receptors on neurons of all types and also the origin of the axons from which they receive glutamate. Learning about neurons, their behaviour and their connections will not by itself solve our problems, but will at least suggest the sort of answers to look for and can be used, often rather decisively, to disprove false theories.

Thanks are due to Patricia Churchland, David Rumelhart and Terry Sejnowski for helpful comments. This work was supported by Kieckhefer Foundation and the System Development Foundation. □

Francis Crick is at The Salk Institute, PO Box 85800, San Diego, California 92138-9216, USA.

GUIDE TO AUTHORS

Please follow these guidelines so that your manuscript may be handled expeditiously.

Nature is an international journal covering all the sciences. Contributors should therefore bear in mind those readers for whom English is a second language and those who work in other fields. Please write clearly and simply, avoiding unnecessary technical terminology. *Nature*'s staff will edit manuscripts to those ends if necessary. Contributors should check their proofs carefully.

Because of the competition for space, many of the papers submitted for publication cannot be accepted. For this reason, and because brevity is a great assistance to readers, papers should be as brief as is consistent with intelligibility. Please note that one printed page of *Nature*, without diagrams or other interruptions of the text, has fewer than 1,300 words.

CATEGORIES OF PAPER

Review Articles survey recent developments in a field. Most are commissioned, but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews Coordinator. Length is negotiable in advance but must not exceed six pages of *Nature*.

Articles are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text or more than six display items (figures and tables) and should not occupy more than five pages of *Nature*.

Articles should be accompanied by a heading of 50 – 80 words written to advertise their contents in general terms, to which editors will pay particular attention. A heading (printed in italic type) is not an abstract and should not usually contain numbers or measurements. The study should be introduced in more detail in the first two or three paragraphs, which should also briefly summarize its results and implications.

Articles may contain a few subheadings of two or three words. The meaning of the text should not depend on the subheadings, whose function is to break up the text and to point to what follows. There should be fewer than 50 references.

Letters to Nature are short reports of outstanding novel findings whose implications are general and important enough to be of interest to those outside the field. Letters should not have more than 1,000 words of text or more than four display items and should not occupy more than two pages of *Nature*. The first paragraph should describe, in not more than 150 words, the origins and chief conclusions of the study. Letters should not have subheadings or more than 30 references.

Commentary articles deal with issues in, or arising from, research that are also of interest to readers outside research. Some are commissioned, most are unsolicited. They are normally between one and four pages of *Nature* in length.

News and Views articles are intended to inform non-specialist readers about a recently published advance. Suggestions should be made to the News and Views Editor. Illustrations are welcome. Proposals for meeting reports should be agreed in advance.

Scientific Correspondence is for the discussion of scientific matters, including contributions published in *Nature*. Priority is given to contributions of less than 500 words and five references. Figures are welcome.

GENERAL

All contributions submitted for publication in *Nature* should conform with these rules.

Manuscripts should be typed, double-spaced, with a good-quality printer, on one side of the paper only. Four copies are required, each accompanied by lettered artwork. Four copies of half-tones should be included. Reference lists, figure legends and so on should be on separate sheets, all of which should be double-spaced and numbered. Copies of relevant manuscripts in press or submitted for publication elsewhere should be included, clearly marked as such. Revised and resubmitted manuscripts should also be clearly marked as such and labelled with their reference numbers.

Titles should say what the paper is about with the minimum of technical terminology and in fewer than 80 characters. Authors should avoid active verbs, numerical values, abbreviations and punctuation and should include one or two key words for indexing purposes.

Artwork should be marked individually and clearly with the author's name, figure number and, when known, manuscript reference number. Original artwork should be unlettered. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are permitted only if the parts are closely related, either experimentally or logically. Suggestions for cover illustrations, with captions, are welcome. Original artwork (and one copy of the manuscript) will be returned when a manuscript cannot be published.

Colour Artwork. A charge of £500 a page is made for 4-colour figures. Inability to meet these costs will not prevent the publication of essential colour figures if the circumstances are explained.

Figure legends must not exceed 300 words and ideally should be much shorter, and should use telegraphic form wherever possible. The figure should be described first, then, briefly, the method. Reference to a method described elsewhere is preferable to a full description. Methods should not be described in detail in the text.

References should be numbered sequentially as they appear in the text, followed by those in tables and finally those in the figure legends. Reference numbers apply only to papers published or in the press, and a different number should be given to each paper cited. All other forms of reference (including unrefereed abstracts) should be included in the text as personal communication, manuscript in preparation or preprint (with number and institution where appropriate). Text should not be included in the reference list. References should be abbreviated according to the *World List of Scientific Periodicals* (fourth edition, Butterworth, London, 1963 – 65). The first and last page numbers should be cited. Reference to books should clearly indicate the publisher and the date and place of publication.

Abbreviations, symbols, units and Greek letters should be identified the first time they are used. Acronyms should be avoided as much as possible and, when used, defined. Editors will shorten words if necessary. In case of doubt, SI units should be used.

Footnotes should not be used except for changed addresses or to identify the corresponding author if different from the first-named.

Acknowledgements should be brief. Grant numbers and contribution numbers are not allowed.

***Submission.** All manuscripts may be submitted to either London or Washington. They should not be addressed to editors by name. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value $5' to prevent the imposition of import duty and value-added tax (VAT).*