Editorial

What is caught in neural nets?

Net: Anything reticulated or decussated at equal distances, with interstices between the intersections.
(Samuel Johnson: Dictionary 1st edition, 1755)
Net: Anything made with interstitial vacuities.
(Samuel Johnson: Dictionary 8th edition, 1808)

We generally think of the brain as a supremely complicated machine—as a biological supercomputer. So, in a sense, brain science is a branch of engineering. I expressed this thirty years ago in a paper called “The brain as an engineering problem” (Gregory 1961). A central point was how difficult—impossible—it is to localize brain function, or say where processes are going on, without knowledge of what the internal functions are in essentially engineering terms. This kind of description overcomes the phrenologist’s fallacy of supposing that behavioural characteristics are directly related to processes in specific regions.

We may say that components, such as nerve cells and switches and condensers, are localized. And we may see their functions as more-or-less localized. But it is impossible to say, for example, where the power of an engine is localized, for many parts and functions contribute to the final result. This makes interpreting brain ablation studies difficult, for brain functions are typically extended over many interacting neural components.

To cite the example I used at that time (Gregory 1958):

“If any changes take place upon removal of part of the brain, the changes are either (a) loss of some feature of behaviour, or diminution or worsening of some skills, or (b) introduction of some new behavioural features. Now it is often argued that if some part of behaviour is lost...the causal mechanism necessary for this behaviour is localized in the affected region. But does this follow? ... We can take an example from radio engineering. If a main smoothing condenser breaks down (shorting the H.T. to earth through a low resistance), the set may stop working, or work in a peculiar manner. The local oscillator may stop, although the rest of the system may continue working normally. Would we then say that the condenser was functionally important for the radio’s oscillator, but not for the rest? If so, we would surely be wrong. Its purpose in the system is to smooth the ripple for the whole system, but it happens that this part of the system is more sensitive to reduction in supply voltage than the rest. Suppose that when the condenser breaks down, the set emits piercing howls. Do we argue that the normal function of the condenser is to inhibit howling? Surely not. The condenser’s abnormally low resistance has changed the system, and the new system exhibits new properties—in this case howling.”

This does bring out the difficulties of establishing and localizing functions even in familiar and well understood interactive systems. It points, too, to how difficult it is to explain phenomena of abnormality or malfunction—for we are faced with a different, new, and even unique system, perhaps beyond available theoretical or practical understanding. It also points to the need for levels of description—with more or less specific localizations of component parts and functions—which it is important to distinguish clearly. The levels of description I suggested thirty years ago, as illustrated in the figure for a simple circuit, were (Gregory 1961):

- **blueprint:** showing the appearance of components—the anatomy;
- **circuit diagram:** showing the functional properties of the components—the physiology;
- **block diagram:** showing the flow of power or information, or the procedures, or operating rules.
This was written before digital computers were commonplace. We were thinking then in analogue terms; especially of interactive neural nets, as in Donald Hebb’s *Organization of Behaviour* (1949) which was much discussed at that time, and the control concepts of cybernetics, which were developed for weapons during the war and were seen as conferring life-like purposeful behaviour on some machines.

The notion of perception operating according to rules was familiar from writings of the Gestalt psychologists. Rules or laws of perceptual organization were described, with examples of dots forming meaningful patterns. Brain interactions were seen as important for meaning (Wertheimer 1922):

“Hence we may say in general that a whole is meaningful when concrete mutual dependency obtains among its parts. The mosaic or associationistic hypothesis is therefore on principle unable to supply any direct approach to the problem of meaning”.

No doubt through the later familiarity with digital computing, the Gestalt rules or laws (Kohler 1920; Wertheimer 1923) are now often called ‘algorithms’. As I shall go on to discuss, I think this is a mistake.

**Rule-based illusions**

An interesting notion, and a significant use of perceptual rules, is to allow present data to be handled from past experience, making it possible to anticipate the future. But when the present is unusual, the guiding rules from the past can lead to systematic errors—illusions. This is bound to happen in some situations, for rules work best when they are broad and general; therefore they cannot apply in all special cases. For vision this seems to be so for the perspective geometry of retinal images, which gives safe rules for seeing typical three-dimensional object shapes from two-dimensional retinal images—though not without problems for pictures and for queer-shaped objects such as the Ames room, chair, and window (Ittleson 1952). Thus also, distortion illusions may be due to following normally sound rules for size-scaling, which mislead in picture perspective (Gregory 1963, 1970, 1980). Their interest for me was, and still is, that these dramatic phenomena occur not because of limitations or failures of components, or of physiological function, but rather at the level of general operating rules. So these phenomena should cast light on some basic psychological principles.

Such rules are so useful they should apply beyond brains to possible seeing machines (Gregory 1967). Some illusions should be much the same for the machine as for us, though its ‘physiology’ is quite different, if its operating rules are similar to ours. In this kind of case explanations of the phenomena must lie at this deep level of description of operating rules, and how they are applied or misapplied.

For valid explanation of an illusion it is necessary to know whether such a phenomenon is due to:

- failure of components;
- inappropriateness of functions;
- inappropriateness of operating rules.

The kinds of explanation are extremely different for each of these levels. So it is very important to classify and explain illusions—or any phenomena—according to appropriate descriptive levels.

Thirty years ago, analogue devices were more common than digital computers, which were in their unreliable and clumsy though very exciting infancy. At that time we did ask: Are the brain’s operating rules carried out by analogues or digitally? Curiously, this distinction seems now often to be ignored, perhaps with disastrous consequences.

The analogue–digital distinction is sometimes made in terms of *continuous* (analogue) and *discontinuous* (digital) processing. But this is not basic: it fails to
capture the essential point. The essential point is that digital computers follow explicitly formulated algorithms; while analogue systems do not have algorithms, and do not go through computational steps. So analogue 'computers' are not computers! They avoid the need for computing, by following analogue pro-formas.

The most familiar example is a graph. If the average height of children is plotted against age, we can read off the average or expected height for any age with no computing. Analogues like this can be built into a machine in many ways: mechanical cams, which may have any form; or physical laws may be used, such as the acceleration of a falling weight or the periodicity of a pendulum; or simple circuits, such as condensers with parallel resistors which may serve as an integrator—again with no steps of computing. This freedom from computing gives analogue systems, even with slow components, great speed; though they do not give exact answers and they are not very flexible, so are used mainly for dedicated special purposes.

The vital point is that, though both analogue and digital devices work with rules, only digital devices use the computing steps of algorithms. Analogue pro-formas may be described with algorithms, but they do not follow algorithms. Analogue systems do not have algorithms, because they do not carry out computations.

Let's take this right outside brains—even beyond the Earth—to the Moon. The Moon orbiting the Earth follows laws of physics, which may be described by algorithms, and can be used to compute future positions. But the Moon does not itself carry out computations to orbit the Earth. It does not compute, or use algorithms—though the astronomer may compute and use algorithms to describe and predict its motion. An artificial satellite may, however, have its orbit set and controlled by a digital computer following algorithms. This is done for course corrections. For contrast, a familiar analogue instrument is a car speedometer: this does not compute \( v = \frac{d}{t} \), though \( v \) is defined as the distance travelled divided by the time taken. It works with a simple analogue system without computing or using any algorithm. No doubt we see velocities (Mather 1990) without computing or using algorithms!

In general: Does the brain follow pro-formas of analogues—or computing rules of algorithms?

David Marr, in his influential book *Vision* (1982), speaks throughout of algorithms of visual processing. Is he assuming the brain to be a digital computer? As Marr himself emphasized and did a great deal to discover, in the early stages of vision there are various kinds of filters. But, surely, it is most implausible to suppose that they work by digital computing. True, digital computers are used by us to simulate and investigate properties of analogue filters; but a simulation is not the same as what is being simulated, and these are essentially different. For it is virtually certain from inspection of the neural 'components' that the filters of visual processing are analogue—and so lack algorithms—though digital computing is employed to simulate their characteristics for research purposes.

Does saying that processes of vision are analogue—though usefully described by digital algorithms—have practical consequences? Yes—for when these distinct though easily confused kinds of systems are pushed to their limits, and beyond, very different things happen. Computer scientists choose algorithms which may not be maximally efficient in favour of more robust though less efficient algorithms which work over a wide range of conditions without breaking down (Sedgewick 1988). But the ways algorithms break down, when they do, is typically different from analogue failures. So there should be errors of perception (illusions and so on) which would distinguish between them, for the functioning of the brain. Adaptations to tilt and curvature may well be a case in point. They strongly suggest modifications of analogue physiological processes rather than breakdown of digital algorithms.
Algorithms of digital systems hold within the limits of their logic: to break down for logical reasons. But this is not so simple, for digital systems may suffer component failure, so nonalgorithmic changes may take over unexpectedly. Thus drugs might change perception and behaviour for digital processes, though without changing the algorithms. Theories of some illusions should be different for analogue or for digital processing; but where it is simply the rules underlying analogue pro-formas or digital algorithms that are inappropriate, the accounts should be the same. This is the deepest kind of explanation we can hope to find. It should transfer to artificial intelligence.

If we look at some more recent accounts and ideas in these terms, what do we find? David Marr (1982) set up his well known levels of description as:

- **the hardware implementation**—the components and how they are wired;
- **the algorithm**—the rules to be followed for digital processing;
- **the computational theory**—the problem to be solved, in terms of the available algorithms.

For David Marr, the most important level is the algorithm. Then comes the computational theory and last the hardware. This is fine for neural (or silicon or wheeled) computers working by following steps of algorithms—but surely it is not appropriate for analogue systems (which much of the brain may be) as they do not have algorithms, or presumably ‘computational theories’. If we are right here, that the analogue–digital distinction is not merely verbal, but is conceptual with empirical consequences, this is a matter of real importance.

We may, indeed, want to reconsider much of what David Marr says. For example we might translate (page 23): “An algorithm by which the transformation may actually be accomplished” (his italics), to something more like: “A rule by which the transformation may actually be accomplished” (my italics). Then the option of analogue or digital with their different implications is left open. To my mind the neutral term ‘rule’ is very useful as it can apply equally to analogue or digital processing.

What could noncomputing brain mechanisms be like? If, again, we go back to earlier ideas, we return to the Cambridge psychologist Kenneth Craik (1914–45) who thought of the brain as having “mental models” representing reality and being thoughts. For Craik they were physically in the brain (The Nature of Explanation, 1943, page 51):

“By a model ... I do not mean some obscure non-physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the process it parallels .... Thus the model need not resemble the real object pictorially; Kelvin’s tide-predictor, which consists of a number of pulleys on levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects—it combines oscillations of various frequencies so as to produce an oscillation which closely resembles in amplitude at each moment the variation in tide level at any place.”

But there are certainly no pulleys or levers in the brain. What the brain has are billions (between 10 and 18 billion in both hemispheres) of nerve cells, of various identifiable kinds, with as many as two thousand connections for each cell. In man there are over half a billion cells in the visual cortex (Blinkov and Glezer 1968). What do they do? Are they simply on–off computer-like switches?

The American neurologist, psychiatrist and philosopher Warren McCulloch (1899–1969) concluded that each brain cell is an elaborate analogue device affected by its many inputs (electrical and chemical) with a subtly regulated threshold for firing. In the early 1940s McCulloch and the logician Walter Pitts saw the brain nerve cells as working cooperatively in small groups forming ‘psychons’, units of thinking (McCulloch and Pitts 1943) which could be analysed in detail. They drew hypothetical circuits for the psychons and these could actually be made with artificial neurons. But, of course, a few artificial psychons could not make a psyche so the general appropriateness of their
Editorial

565

notions was hard to test. The cell activity in the nets was not supposed to vary linearly, but to fall off gradually with use associated with learning, or the net would soon be saturated. This nonlinear feature made mathematical analysis and precise predictions for large nets very difficult.

Hebb (1949) considered much larger nets, randomly interconnected, forming internal patterns of activity modelling objects and thoughts and memories—each cell becoming more active as it is stimulated more often—to build up very large cortical patterns. This is the prototype for all the recent PDP (parallel distributed processing) theories.

Attempts were made early on to construct seeing machines with artificial neural nets. Most famous was Frank Rosenblatt’s Perceptron (Rosenblatt 1962). With excitatory and inhibitory connections (synapses) it could begin to generalize patterns. There were two layers of artificial neurons, generally with every input (retinal receptor) connected to every output. With the cells adapting to frequency of use this system could give some generalization of patterns. Marvin Minsky and Seymour Pappert (1969) showed, however, that there must be severe limitations in Rosenblatt’s Perceptrons, and partly from their valid criticisms the idea was dropped. Interest declined in this approach also because of the evident power and flexibility of the digital computer; though von Neumann, its inventor, did himself realize the potential power of interactive nets for some uses. There are many exciting papers written at that time with ideas that should have flowered, for example the conference Neural Networks edited by the Italian pioneer Caianiello (1968). If interest in Perceptrons had not died, it might have been realized much sooner that adding more layers between inputs and outputs makes a fundamental difference. These ‘hidden units’ allow inner patterns to develop, which are not driven by the input or closely related to the output. They are hidden and secret much as mind is hidden and secret. They have remarkable powers to abstract and learn and discover generalizations. Perhaps, suggestively, they need periods of rest to sort themselves out and apparently they dream!

The new nets are inspired by the work of John Hopfield (1982), Geoffrey Hinton and Terrence Sejnowski (1986), and now many others. Michael Arbib, who was a student of Warren McCulloch, has developed interesting ideas (Arbib 1989). Ivor Aleksander worked for years developing the impressive WISARD (Wilkie, Stonham, and Aleksander’s Recognition Device), which, with other recent work, is clearly described in Aleksander and Morton (1990). Terrence Sejnowski has devised a net that can learn to read English and synthesize its own speech very much as children do, starting with random babbling. His artificial net is faster than babies at learning from examples without algorithmic programs—though one does not know how sophisticated it can become. This is a general query: many nets work well for small problems, but rapidly become inadequate with increasing number of alternative possibilities. How best to teach artificial neural nets presents unsolved problems. But here may lie some helpful hints and perhaps significant research for human education.

The essential point of PDP nets is that analogue interactive systems can learn from successive presentations of faces, or of differently written letters or sounds of words or whatever—to generalize and recognize new objects of the same class. And they may devise the classes. Unlike digital devices, they go on working though a large proportion of their components is destroyed. And they can cope with partial or mangled input signals, as we can. Is this where the answer lies for how the brain works? Or are brain processes digital programs (which might be transferred from one brain to another, as we swap floppy computer discs)? The snag of neural nets is that their nonlinear interactive functioning is very hard to analyse. It is a bizarre thought that, though we may make effective and brain-like artificial neural nets, they may remain black to our understanding. In other words, we may fail to devise algorithms to describe them.
Implications for AI

The analogue–digital distinction has interesting implications for artificial intelligence. The strong claim is that a machine can be build that will display all the intelligence and have the consciousness of the higher organisms, especially humans. Recent philosophers of AI have considered digital computers rather than analogue devices. This may be a mistake.

If the “strong AI” notion that a machine can have a mind is true, a mind-full machine as it has been said could “be made of old beer cans”. On a digital algorithm view this would be so if the beer cans can count. On an analogue view the ‘old beer cans’ would need to have far more subtle and elaborate characteristics—for the functional units are much larger than the simple switching steps of digital systems, as analogue pro-formas embody wide-ranging functions. An entire range of input–output values might be carried on a single pro-forma.

If we apply these considerations of analogue and digital to AI philosophy, we may want to amend some recent criticisms. The American philosopher John Searle has mounted a sustained and influential attack on digitally conceived AI. One of his points is (Searle 1984, page 36):

“The question isn’t: ‘Can a machine think?’ or: ‘Can an artifact think?’ The question is: ‘Can a digital computer think?’ But again we have to be very careful how we interpret the question. From a mathematical point of view, anything whatever can be described as if it were a digital computer. And that’s because it can be described as instantiating or implementing a computer program. In an utterly trivial sense, the pen that is on the desk in front of me can be described as a digital computer ‘Stay there’. Of course our brains are digital computers, since they implement any number of computer programs.”

Really? Anything (following Turing 1950) can be described as computer programs. But this does not mean they are computer programs. Searle continues:

“And of course our brains can think. So once again, there is a trivial answer to the question. But that wasn’t really the question we were trying to ask. The question we wanted to ask is this: ‘Can a digital computer, as defined, think?’ That is to say: ‘Is instantiating or implementing the right computer program with the right inputs and outputs sufficient for, or constitutive of, thinking?’ And to this question, unlike its predecessors, the answer is clearly ‘no’. And it is ‘no’ for the reason that we have spelled out, namely, the computer program is defined purely syntactically. But thinking is more than just a matter of manipulating symbols, it involves meaningful semantic contents. These semantic contents are what we mean by ‘meaning’.”

John Searle’s point is that digital computer programs have syntax (grammar) but not semantics (meaning) and so cannot be genuinely intelligent. Searle holds, essentially, that computers may be able to simulate human behaviour but: “If we are talking about having mental states, having a mind, all of these simulations are simply irrelevant”. So, he rejects strong AI. Some computer scientists hold, however, that programs can have semantics: after all, computers can tell us which planes have got free seats.

We might well ask how brain programs (and Searle is considering the brain as a digital computer) can have semantics if man-made computers can’t. Searle’s claim is that:

“Mental states are biological phenomena. Consciousness, intentionality, subjectivity and mental causation are all part of our biological life history, along with growth, reproduction, the secretion of bile, and digestion.”

This is a surprisingly vitalist creed. For over a century the story of biology has been the gradual rejection of special substances for life; since ‘organic’ chemistry became associated with carbon, rather than with special substances for life, following the production of ‘organic’ urea by heating ‘inorganic’ ammonium cyanate, by Friedrich
Wohler in 1828. Searle appears to be backtracking, seeking some special uniquely biological life substance for mind.

If the brain is not carrying out algorithms, but is rather an interacting net or some such analogue system without algorithms, does this affect Searle's position? Presumably he would say that we have moved the goal posts of his position—that he is only talking about algorithms—but he might maintain that only biological substance can provide analogue pro-formas of whatever for intelligence. How this could be a philosophical, rather than an alleged empirical conclusion beats me. The fact is we do not know any limits of this kind. How can we possibly be sure, on any philosophical grounds, that an artificial net or digital computer programs properly carried out could not embody the full attributes of biological mind? Surely this is a question only experiment can answer.

A more recent somewhat related attack on AI is made by the highly distinguished mathematician, cosmologist and inventor of impossible objects and much else, Roger Penrose, in his book *The Emperor's New Mind* (1989), which has already stimulated much discussion. It has very interesting things to say, including fascinating insights on the nature and importance of algorithms. Penrose also equates AI accounts of brain function with carrying out algorithms. Does he (as I have suggested for David Marr and for Searle) conflate the use of algorithms for doing (implementing) and describing (simulating)? He gives algorithms and computing a wide brief. Penrose speaks of quanta, and other physical entities and events, as 'computing'. Evidently he thinks of physical laws as algorithms, somehow given by nature. This is very different from the position I have adopted; which is that algorithms are not in nature, except in systems that describe—brains and computers—using rules of mathematics and logic.

This position that natural laws are algorithms surprises me. I would have thought we use algorithms to describe and compute from physical laws—not that the laws themselves are algorithms. The distinction is clearly important: let the reader decide!

Roger Penrose bravely discusses consciousness, suggesting that the key may lie in developments of quantum mechanics. He rejects algorithms as vehicles for intuitive thinking and consciousness, (including for mathematicians' intuitions) preferring quantum properties, perhaps not yet discovered, of matter. But we want to avoid the conclusion that all matter (tables, chairs, wood) is intelligent and conscious. So there must be something special about brains. If, though, these supposed properties of matter can be specially used in brains why not also in man-made brain-like machines—for knowing and feeling much as we do?

Richard L Gregory

**References**

- Aleksander I, Morton H, 1990 *Neural Computing* (London: Chapman and Hall)
- Arbib M A, 1989 *The Metaphysical Brain 2: Neural Networks and Beyond* (New York: John Wiley)
- Blinkov S M, Glezer I, 1968 *The Human Brain in Figures and Tables: A Quantitative Handbook* (New York: Basic Books)
- Caianiello E R, 1968 *Neural Nets* (Berlin: Springer)
- Craik K, 1943 *The Nature of Explanation* (Cambridge: Cambridge University Press)
- Gregory R L, 1958 "Models and the localization of function in the central nervous system" in *Mechanization of Thought Processes* Volume 2, National Physical Laboratory Symposium No 10 (London: HMSO) [reprinted in: Gregory R L, 1974 *Concepts and Mechanisms of Perception* (London: Duckworth) pp 537–542]
- Gregory R L, 1961 "The brain as an engineering problem" in *Current Problems in Behaviour* Eds W H Thorpe, O L Zangwill (London: Methuen) [reprinted in: Gregory R L, 1974 *Concepts and Mechanisms of Perception* (London: Duckworth) pp 547–565]
- Gregory R L, 1963 "Distortion of visual space as inappropriate constancy scaling" *Nature (London)* 199 678–691
Gregory R L, 1967 “Will seeing machines have illusions?” in Machine Intelligence IV Eds N L Collins, D Michie (London: Oliver and Boyd) 169 – 177
Gregory R L, 1970 The Intelligent Eye (London: Weidenfeld and Nicolson)
Gregory R L, 1980 “Perceptions as hypotheses” Philosophical Transactions of the Royal Society of London B 290 181 – 197
Hebb D O, 1949 Organization of Behaviour (New York: John Wiley)
Hinton G E, Sejnowski T J, 1986 “Learning and relearning in Boltzmann machines” in Parallel Distributed Processing. Explorations in the Microstructure of Cognition Eds J L McClelland, D E Rumelhart, volume 1 (Cambridge, MA: MIT Press) pp 282 – 317
Hopfield J J, 1982 “Neural networks and physical systems with emergent collective properties” Proceedings of the National Academy of Sciences of the USA 79 2554 – 2558
Ittleson W H, 1952 The Ames Demonstrations in Perception (Princeton, NJ: Princeton University Press)
Kohler W, 1920 “Physical Gestalten” reprinted in A Source Book of Gestalt Psychology Ed. W D Ellis (1938, Henley-on-Thames, Oxon: Routledge and Kegan Paul) pp 17 – 54
McCulloch W S, Pitts W, 1943 Bulletin of Mathematical Biophysics 5 115 – 133
Marr D, 1982 Vision (San Francisco, CA: Freeman)
Mather G, 1990 “Computational modelling of motion detectors: responses to two-frame displays” Spatial Vision 5 (1) 1 – 14
Minsky M, Pappert S, 1969 Perceptrons (Cambridge, MA: MIT Press)
Penrose R, 1989 The Emperor’s New Mind (Oxford: Oxford University Press)
Rosenblatt F, 1962 Principles of Neurodynamics (East Lancing, MI: Spartan)
Searle J, 1984 Minds, Brain and Science (London: BBC)
Sedgewick R, 1988 Algorithms (Reading, MA: Addison Wesley)
Turing A M, 1950 “Computing machinery and intelligence” Mind 59 433 – 460
Wertheimer M, 1922 “Gestalt theory” reprinted in A Source Book of Gestalt Psychology Ed. W D Ellis (1938, Henley-on-Thames, Oxon: Routledge and Kegan Paul) pp 1 – 11
Wertheimer M, 1923 “Laws of organization in perceptual forms” reprinted in A Source Book of Gestalt Psychology Ed. W D Ellis (1938, Henley-on-Thames, Oxon: Routledge and Kegan Paul) pp 71 – 88