THE SECOND LECTURE ON EVOLUTION, CONSCIOUSNESS AND CONSCIOUS EVOLUTION

A systems approach to consciousness

PROFESSOR AARON SLOMAN
School of Computer Science, University of Birmingham

Delivered to the Society on 26 February 1996, with Professor Igor Aleksander, Department of Electrical & Electronic Engineering, Imperial College of Science, Technology & Medicine, in the Chair★

THE CHAIRMAN: This lecture is the second of three in a series on evolution, consciousness and conscious evolution.

I asked Professor Sloman to describe himself and he did so as a failed mathematician and renegade philosopher masquerading as a computer scientist.

The truth, of course, is that he has had a brilliant career, starting with a degree in physics and mathematics, pure and applied (First class) from the University of Capetown, followed by a DPhil in Philosophy from Oxford University. He is one of the leading lights in the field of artificial intelligence, where I would describe him as one of the few thinkers.

I am embarrassed to be writing about consciousness because my impression is that nearly everything written about it, even by distinguished scientists and philosophers, is mostly rubbish and will generally be seen to be rubbish at some time in the future, perhaps two hundred years from now. This means my own work is probably also rubbish.

However, I shall outline some ways in which we can hope to make progress through collaborative, multi-disciplinary research. By doing this work we are sure to learn *something*.

First, I need to distinguish three different types of questions:

- empirical,
- design
- conceptual.

Design questions are at the heart of our problems.

EMPIRICAL QUESTIONS

Empirical questions are about what exists in the world, what is correlated with what and what happens if you poke it, give it electric shocks or whatever. Much empirical research is interesting and valuable, until the researchers start claiming they have discovered something important about consciousness, as if consciousness were a well defined topic and a proper object of research, like influenza, magnetism, or carbon. I believe there is no unique, clearly identified concept. Therefore much of what is written about how experimental results relate to consciousness is ambiguous and muddled, even if the experiments themselves are interesting and tell us useful things about human or animal brains, such as which bits of the brain react to which sorts of sensory stimulation.

DESIGN QUESTIONS

Design questions are about what kind of mechanisms and capabilities can be put together in various ways in order to achieve various richer kinds of capabilities. Many old philosophical questions are now best discussed in design mode, which is why I am in a School of Computer Science, although I am actually a philosopher.

★ *This text was prepared for the Journal by Professor Sloman. A full transcript of the lecture is available from the Lectures Office: audio tape £5, paper £3, IBM-compatible floppy disk £2.50 (all inc. VAT).*

Engineers, whether concerned with large or small scale, with physical or abstract software structures, all have to think about designs. However, our understanding of the space of possible designs is still primitive. Unfortunately, many people whose knowledge about designs for working systems is shallow, presume they have a deep understanding and can make pronouncements about what certain classes of designs can and cannot do, e.g. pronouncements about the likelihood, or the impossibility, of computer-based systems having consciousness.

CONCEPTUAL QUESTIONS

Conceptual questions are commonplace both in philosophy (e.g. attempts to clarify what we mean by words like 'good', 'true', 'possible', 'experience') and in deep science. Examples are Einstein's analysis of the concept of 'simultaneity', biologists' attempts to clarify concepts like 'function', 'gene', 'species' and the struggles of modern physicists to clarify concepts used in quantum mechanics.

Part of my criticism of much writing on consciousness is that it ignores deep and difficult conceptual questions. Authors assume first-hand knowledge about consciousness is enough. That's as unjustified as claiming that long before Einstein we all knew what simultaneity was because we had first hand knowledge about it.

A mind, or a brain, is, above all, a control system. Many concepts, such as 'consciousness', 'pain', 'emotion', 'personality' cannot be clarified except via study of possible designs for complex information processing systems: when we know something about the architecture of a system we can use that as a basis for generating and classifying concepts describing the kinds of states that system can be in, just as theories about the architecture of matter led to a new set of concepts about kinds of stuff, e.g. in the periodic table. We have yet to develop an architecture-based periodic table for the mind.

Many who think they know about consciousness from direct acquaintance claim there is a binary division (a dichotomy) between things that do and do not have consciousness, or a binary division between states in which we are conscious and states when we are not, or a binary division between those things of which we are conscious at a time and those of which we are not.

That illusion tempts us to ask questions such as 'Which animals have consciousness and which don't?' How did 'it' evolve? Does 'it' have a biological function? Is 'it' reducible to physics? Could a robot have 'it'? Could software running in a computer have 'it'? Could there be a machine which is a 'zombie', with all the appearance and behaviour normally associated with consciousness but totally lacking 'it'? All these (and many other) questions become vacuous if the concept of consciousness turns out to be a muddled collection of different concepts.

The enormous diversity among living things (e.g. sunflowers, carnivorous plants, amoebas, rats) reveals no obvious place to draw the line. Similarly, when a foetus develops, it is at first simply a cell and then it starts dividing. Eventually it pops out and yells and, eventually, gives lectures on philosophy. Is there a time at which it switches from something without consciousness to something that not only has 'it' but talks about 'it'?

Later, the individual may have a degenerative brain disease, suffering slow degradation until what is left is just a 'vegetable' (or rather a piece of meat). At what point does consciousness cease?

Here's an example illustrating the ambiguity regarding what is in 'it'. Many people will have seen this phrase, often shown displayed inside a triangle:

A
BIRD
IN THE
THE HAND

People can stare at it and not see anything wrong, even when told there is something wrong (e.g. a subset of those reading this). Yet, if they are told to shut their eyes and then asked certain questions (such as how many words there are, or where the 'the' is), they sometimes realise they did see something wrong. They saw it but they somehow did not pay attention to it. Were they somehow conscious of it and not conscious of it at the same time?

Susan Greenfield mentioned several of these problems. Her response was that there is no dichotomy but a continuum: i.e. there is no sharp division between things with and without consciousness, or between what you are and are not conscious of because it is all a matter of degree.

I think this is partly correct but also dangerously misleading: because the idea of a smooth continuum here is wrong. What many people don't realise is that a

continuum is not the only alternative to a single major discontinuity.

There could be a large number of different discontinuities. If so, we can embark on the exciting research programme of exploring all the myriad divisions, instead of seeking the mythical unique boundary line, or wallowing in the intellectually un-challenging idea that there are no divisions.

So, while there is no dichotomy (no sharp distinction between things with and without consciousness or between the things we are and are not conscious of) it is also a mistake to think we can avoid the problems by thinking of consciousness as just a matter of degree.

Of course there are lots of differences of degree between animals but accepting that leaves out a much more important truth, namely that we are talking about a *cluster* concept. To say 'consciousness' is a cluster concept is to say it is neither something unitary that is present or absent, nor something smoothly varying in degree: rather it involves a large cluster of recombinable capabilities which can be present or absent in different combinations. They are not simply things that vary in quantity. They are different in their function, structure, origins and ways of going wrong. Those are differences in kind, not in degree.

Different subsets of the cluster can occur in different organisms, in different sorts of machines, different people (we are not all exactly the same – we have different kinds of capabilities). Even in the same person, at different times, the collection of capabilities may change: between infancy, childhood and adulthood and in senile dementia, brain injury and so on.

When I say consciousness is a cluster concept involving a collection of properties – a, b, c, d, e, etc. – I am not saying consciousness is some logical combination of them, such as a disjunction, conjunction of disjunctions, or disjunction of conjunctions or whatever.

The concept has a kind of indeterminacy as to what is and is not required in its instances, partly because our grasp of what is possible is too limited for us to have clear notions about how to divide things up and partly because our previous history has not forced us to agree on such criteria. So, cluster concepts can be indeterminate and consciousness is no exception.

What sorts of capabilities are involved in this concept? I have space only for a few examples. Readers may be able to produce many more.

Humans have many kinds of perceptual capabilities. Being able to recognise something as an 'a', a 'b' or a particular word, is different from interpreting it as having a meaning, i.e. recognising objects is different from interpreting them. Another sort of interpretative capability involves the ability to use vision to derive a rich three-dimensional structure from two-dimensional patterns of stimulation on the retina.

Yet another profound capability is the ability not only to perceive structures but also possibilities and constraints on possibilities, for instance, seeing that a window catch admits certain possible states and some of those states constrain motion of the window. Not all people can see the same possibilities and constraints. Cultures can vary, just as animals do.

There are also different sorts of memory capabilities. There are short-term memories, like the ability to store and repeat the numbers 7, 3, 5, 2, 8, 2. The ability to remember some things for longer may enable an animal to recall that it has a nest a long way from its current location and remember how to get back there.

Humans also have an important ability to remember things concerned with the future, e.g. we often need to contemplate several possible future actions and events and move back and forth between them before deciding. Here, short term memory is not concerned with remembering facts but remembering what might be.

There are also many kinds of learning and of motivational processes (feeling hungry, thirsty, sleepy, tired, hot, cold, sexy or wanting to solve a mathematical or philosophical problem). Humans at least, have what we call moral feelings or can feel indignation about what is being done to something else. How many other animals can do that?

Another class of capabilities involves self monitoring and self control. You can not only look at things around you and do things but you can also attend to what you are looking at, think about what you are paying attention to and notice aspects of one's experience, such as how what one sees varies as one moves, without which much art would be impossible.

Sometimes we control ourselves, e.g. deciding to cease dwelling on yesterday's humiliating experience and instead pay attention to an important task; sometimes we fail. So, we also have the ability to lose control of our thought processes, at least partly. How many other animals can monitor and control thought processes and partially lose control? Those that cannot are incapable of certain sorts of emotional states, which I call 'perturbant'.

These hurried and shallow comments list only a tiny

fragment of the vast collection of capabilities that make up what it is to be a human being and which are all aspects of our having experiences and being conscious.

Different sub-collections may be present in different people, in different organisms, the same person at different times and maybe in machines. Asking which ones constitute consciousness is pointless: the concept is too indeterminate for there to be any answer. In any case, there are far more important questions concerning the study of all these different capabilities, the mechanisms underlying them, how they evolved, which ones can be implemented in computers and so on.

We need to study 'design space': the space of possible designs for systems, with different combinations of capabilities. This is linked to another kind of space, 'niche space', which is partly what biologists study when they talk about the niche of an organism and partly what engineers study when they talk about the requirements against which a design is to be evaluated. Niche space and design space are both complex structures and the mappings between them are complex. What are the dynamics of these spaces: what sorts of trajectories are possible within them? Which sorts can occur in individual development? (As far as I know, it is not physically or chemically possible for a human egg to evolve into a giraffe, no matter how its environment is manipulated.) Which sorts of trajectories require evolution across generations involving many individuals? Which can occur in a laboratory?

Different combinations of capabilities correspond to different designs. Designing an overhead projector involves combining physical capabilities and some cognitive ones (it should be easy to find the controls). Designing a piece of software involves combining more abstract abilities: for instance, to calculate, reformat a document, check spelling or grammar, solve puzzles. Some designs inherently involve a particular physical implementation (e.g. a violin) whereas others admit considerable variety in their implementation (e.g. a particular type of word processor may run on computers with different physical structures).

Evolution also produces designs – even though there is no designer or engineer, just natural selection. The notion of a design is an abstraction that has nothing to do with how the design was produced, or whether any agent intends it to serve any purposes. In that sense the design of a bee fits the requirement to pollinate plants.

I see 'Artificial Intelligence' as a misnomer for a field involving increasingly deep and broad study of regions of design space and regions of niche space, including both natural and artificial systems and including both those that already exist and those that might exist. Cognitive science is the subset of AI that studies the region of human and animal designs. We still do not know much about the overall structure of design space, nor which kinds of mechanisms are needed as substructures for particular sorts of designs. Nor do we know much about the classes of designs that are relevant to characteristically human capabilities.

So, pontificating about whether machines can or cannot be conscious, or about whether consciousness does or does not need quantum gravity engines, seems to me to be just silly. There is far too much research still to be done and then, instead of asking a few big but empty and largely unanswerable questions about consciousness, we can ask, and perhaps answer, a large number of smaller, fascinating questions about all the myriad components of the cluster.

We can study different ways in which transitions can occur from one collection of capabilities to another. Some involve change within an individual, some may not be capable of occurring without an evolutionary process involving generations of individuals. Perhaps an evolutionary process could produce human beings with a strong ability to detect magnetic fields.

Understanding the characteristically human cluster of capabilities requires good understanding of how these are composed of layers with different evolutionary histories.

For example, one kind of agent involves a purely reactive control mechanism. Information is acquired through external sensors and internal monitors and propagates through and around the system and out to effectors of various kinds. Everything happens in parallel because there are dedicated coexisting circuits.

I presume there are many organisms like that and I believe older, more primitive parts of the human brain are like that.

Another architectural layer, a 'management' mechanism, is able to consider options in advance, evaluate and select between them, including constructing new forms of action, for instance, making a new plan. This sort of mechanism, instead of having a lot of hardware dedicated to parallel activities, may have a common store of temporary memory for constructing these 'advance' representations and a common associative memory used to guide the construction evaluation and selection. For a variety of reasons, such a system,

though implemented in intrinsically parallel mechanisms, is intrinsically *serial* in its functionality. This means it is a powerful resource that has to be shared between different goals and needs. Some of the mechanisms for controlling the resource allocation and directing it towards self-improvement could include complex filters and mechanisms involved in what we loosely describe as reward and punishment. I presume that only the more sophisticated animals have these abstract management mechanisms.

With colleagues at Birmingham I have been exploring the notion of a third level of architecture which, in evolutionary terms, would be even more recent and possibly a lot more rare. This involves a 'meta-management' system which can monitor the strategies and behaviour of the management system (and possibly also the meta-management system) and take corrective action when the individual decisions do not seem to be producing an overall state that is valued highly. An example from human life might be noticing that one is switching attention between tasks too frequently, with consequent loss of efficiency. Corrective action might involve deciding to ignore interrupts and new motives for a time. Such control at the meta-management level is not perfect: we sometimes decide, and want, to think about X but are continually drawn back to thinking about Y.

These three layers in the proposed architecture, the automatic reactive processing layer, the management layer which can construct and contemplate options in advance and the metamanagement layer which can monitor, evaluate and redirect high level internal actions, may each add new kinds of capabilities compared with earlier, simpler systems. Trying to treat them as part of a continuum is therefore seriously misleading. Moreover, within each layer there can be a large collection of distinct capabilities, which can be present or absent in different combinations, adding to the diversity of designs and discontinuities in design space.

Some of these, associated with a rich perceptual architecture, can, I believe, account for the phenomena that make many people wish to talk about 'qualia' and 'subjective experience', for these are viewer-dependent aspects of perceptual states. The ability to pay attention to our own experience is different from the ability to have the experience (i.e. the ability to see what is out there). Being able to attend to the experience can be useful for some purposes (drawing things, telling others how things look to us, etc.) I do not know how many other animals which have sensory experiences also have the ability to attend to them, to report them, to compare them.

I'll now return to the conceptual questions. Each design for an architecture determines (together with the environment) a variety of states and processes that the architecture can support. This produces a family of concepts: 'theory based' concepts, grounded in the architecture. These theory based concepts can elaborate and extend common sense concepts. (In my examples I have tried to indicate links with common sense concepts.)

I anticipate such a process of conceptual evolution in philosophy, psychology and biology, to some extent paralleling the conceptual evolution in physics and chemistry that grew out of a new hypothesised architecture for matter.

These new concepts enable us to ask new questions. Instead of asking which animals are conscious and which are not, we might, for instance, talk about 75 different kinds of consciousness and ask which animals have this kind, which have that kind and when do we have this kind and which kind do we lose when we have Alzheimer's disease and which subset do we have when we are very young and so on.

We should not explore only one architecture, or build one type of robot but think about varieties of architectures and trajectories in design space and niche space. Only in terms of similarities and differences can we understand any particular architecture.

These problems are hard and progress will be slow but the prizes to be won are high.

If we have an architecture-based understanding, we can get a much deeper understanding of how a human mind normally works and also how it might go wrong: the more complex the architecture, the more ways it can go wrong. By applying all these ideas, we should be able to help therapy, counselling and education. Most educational theories are based on no understanding of what kind of mechanism we are talking about, just a lot of hunches and rules of thumb.

We can also generate new empirical questions, expressed using new concepts.

We may find most high level human-like architectures can be implemented on a number of different sorts of low level mechanisms – for instance, computer-based mechanisms. I do not know: we might have to invent new mechanisms. We have been trying

to explore what computers can do for only 40 or 50 years. A comprehensive overview may take hundreds of years. We may discover all sorts of new mechanisms and find that if you get the global architecture right, you can get the right kind of system, with several implementation mechanisms with sufficient richness and structural variability.

I expect that the systems approach will show that *architecture dominates mechanism.*

DISCUSSION

THE CHAIRMAN: How can one talk about consciousness without including the mechanisms that have to evolve over a lifetime?

THE LECTURER: You may be falling into the trap of thinking that consciousness is a thing that is or isn't there. I would rather look at what happens with things that can be preassembled and with other things which are not preassembled but have to do a lot of development on their own by interacting with their environment. We might also look at things which can't be designed but which have to evolve. We can compare these things and produce a vocabulary for describing them and perhaps forget about the question of which of them fit some old, obscure notion of consciousness. Perhaps we will find different forms of consciousness.

COLIN TULLY (Consultant, Colin Tully Associates): What are your views on the learning organisation and on self-organisation?

THE LECTURER: I would expect organisations not to have the same kind of coherence as is found in the system I was talking about, or in a system a biological individual would normally have. On the other hand, an ant colony, made up of lots of little agents working together in quite complicated ways, seems to have enormous coherence. We talk about the spirit of the age and the will of the people. We might hope to make these ideas more precise and richer in detail.

I'm sure there's lots of scope for self-organisation of different kinds in the architectures I was talking about. Some of the things that go on in neural nets, which are self-organising, could be important. I'm told the way the visual system develops comes not from DNA but from the system organising itself under the stimulus of incoming signals. There may be other kinds of self-organisation in the more abstract symbolic process that goes on in the higher levels and yet another in a society or organisation. There's an awful lot to explore and study.

GEORGE FINCH (Partner, Finch Macintosh Architects): Would you comment on the importance of verbal language, and therefore logic, in our consciousness?

THE LECTURER: Some control states have structures which are close to what we might think of as linguistic; there are different components which can be separated and recombined in various ways. Others are more abstract. Different kinds of mathematics often involve new formalisms, new types of symbols and new ways of manipulating them. We also have things like pictures and maps and there may be more that we haven't yet thought about.

SUSAN REED (Adviser, The Empowerment Way): Could there be an extra means of communications through senses other than those we're familiar with?

THE LECTURER: I have a completely open mind about what kinds of new forces, modes of communication and modes of transmission will be discovered. Things we take for granted would have been thought of as magic in the past. I will wait and see the evidence for new means of communication.

JOHN FARAGO (Learning Individuals in Learning Organisations): I'm fascinated by the ability of small groups in conversation to gain heightened awareness and improve their learning.

THE LECTURER: It's also fascinating how often social interactions get in the way of people developing their potential. You get factions emerging. As you can see, I've had a lot of experience in universities.

MARTIN C.B. SMITH (Head of Mobile Robots Research Unit, University of East London): I think machine consciousness is relatively easy but machine intelligence is relatively hard. Could you comment?

THE LECTURER: I wouldn't make that kind of distinction. Everything I've said about consciousness I would say about intelligence. It's not a thing you either have or haven't got.

NICK ANDERSON (Physics Teacher): Relating to traditional language, we have meaning attached to the information we process. How can you build a machine in which there is meaning attached to the things it is doing?

THE LECTURER: Some of the subsets involved in our understanding of language are already there in fragmentary form in some artificial intelligence programs. It's possible to make computers take in sentences and do some sort of structural analysis which is vaguely like what a linguist might say we do. That is not understanding but it may be part of it. Meaning also involves relating the structure to something else; computers can do that. They can also build abstract virtual structures and make use of them as a way of representing other things.

Rudimentary forms of reference are already there but if we want to talk about human ability to understand, we have to look at all the different facets of human understanding, which is composed of a large number of different capabilities. I have no reason to believe we have got something that can't be done by suitably designed machines but we have to have an open mind.

JACK STILES: I can't imagine consciousness without sensory input. What are your views?

THE LECTURER: I suspect I'm the only person in the world who would disagree with you. I think you could have a machine which operates only on internal structures. I don't see why it should be impossible to have something which merely exists in the world of mathematics and doesn't have sensors which interact with anything else.

SIR JAMES GOWANS: Being conscious of something depends on a database from past experience. A newborn child would not have this. To what extent does consciousness depend on memory?

THE LECTURER: Human adult consciousness of things that are perceived is deeply imbued with what we know about the world. Concerning a newborn child, my hunch is that evolution has done an enormous amount to prepare us for opening our eyes and starting to learn about the world so the newborn child's apparatus is not a complete blank. There is indirect evidence for this in other mammals: a newborn deer has to be able to get up and run with the herd or find its way to where the milk is. My guess is that there is a lot of knowledge there but it's knowledge acquired not by the individual but by the species.

EDWIN H. COX: In termite mounds, none of the termites have a clue about what they are doing yet we attribute design and structure to the result. So you have a totally reactive process which results in something that looks planned and deliberate. The same seems to happen in the human foetus, when cells behave in different ways without having a clue about what they are doing. Although there is a design, it's not the cells' design. I don't see how consciousness can link in to models though it does link in to processes.

ANDREW STONE (Managing Director, Marks & Spencer): I come from a large organisation which works like a termite colony. One would suppose that consciousness resides at a high level in this organisation but there are other groups who operate from different consciousnesses. The branch in Croydon might be operating with one consciousness and the board of directors with a different consciousness. Any one may come to the fore at any time.

THE LECTURER: You can't necessarily tell from the behaviour of a system what kind of architecture is producing what the system does and achieves. In principle (but not in practice), you could have a computer programmer work out all the possible sequences of inputs and all the possible outputs and the machine might look as if it was doing something intelligent, whereas in fact all the intelligence was in the programmer. There may be something like that in termite colonies but with an evolutionary process instead of a human programmer.

C.J. BROWNE (Research Assistant, Imperial College): Would you agree that symbols in a processing system have to be grounded in some sort of sensory experience so they not only stand for items in the world but represent them in a meaningful way?

THE LECTURER: The whole notion of symbol grounding is confused. It's based on a notion that meanings have to be built up from elements derived from sensory experience, whereas I think of meanings as aspects of complete structures which work in some way. Objects have powers and potentialities as well as structure. A glass has properties of shape and size; there are also various things you can do with it and different ways you can break it. There is something in the glass and the environment that makes the possibilities possible and links them.

The same can be said of information structures. If you see algebraic symbols as marks on paper, they have no powers but when they are implemented in a symbol-manipulating system, something happens. Implemented symbol systems have a lot of powers which might surprise you.