The Mystery of Consciousness

A short fantasy

The year is 1954.

Alan Turing, is returning home from work at the University of Manchester where he is using the recently installed Ferranti Mark 1 computer to further his researches on morphogenesis and other matters. This behemoth of a machine with 4,000 valves, 2,500 capacitors, 15,000 resistors, 100,000 soldered joints and 6 miles of wire boasts a huge 5120 bit random access CRT memory, 72kbytes of magnetic drum storage and can carry out over 800 additions every second; but for Turing this is not nearly enough. He dreams of the day when a computer can play chess as well as he can and can even fool us into thinking that it might be conscious. After all, isn't the human brain just a computer with nerves instead of valves?

“A parcel arrived for you this morning” calls out his cleaner as he closes the front door of his new house in Wimslow. “Someone has been having a right joke! It says on it 'A present from the future'”

It was as she had said. A large cardboard box with his name and address on it and the stated words scrawled across with some kind of painting pen was lying on the table in the hallway.

Later that evening he opened the box and what he found inside astounded him – a beautifully finished silver plastic box the size of a small file which opened up to reveal a keyboard and a shiny black screen. On pressing what appeared to be a power button the machine emitted a quiet whirring sound and the screen sprang to life, glowing in wonderful colour. Alan tried pressing the keys of the keyboard but although nothing much seemed to happen it didn't take more than a few minutes of experimenting for him to get the hang of the sensitive finger pad and the movable arrow. Within the hour he had discovered the 'Games' folder and was playing chess with the machine – and finding it more difficult to beat than he had ever imagined possible.

That night his mind was in turmoil. Where had this machine come from? What exactly could it do? And above all – how does it work?

The next day he discovered an invoice in the box. Apparently the machine cost £299 (more than twice Alan's annual salary at the time) and was purchased in 2013. There was also a small note attached saying 'Hope you find this interesting. It cost me a small fortune to send.' and was signed Bill G.

Alan did not appear back at work for a week during which time he had sussed out most of the machine's capabilities. Of especial interest to him was a program called 'Fortran' which enabled him to write mathematical algorithms and one of the first things he did was to calculate the highest common factor of 2^{18}, a task which took the

first programmable computer (the Manchester 'Baby') 52 minutes to perform in 1948.

Soon, Alan's curiosity began to nag at him incessantly. How does this thing work? Is it a brain? What is it made of? The only screws he could find on the outside gave him access to what was obviously a battery pack and also to a metal box the size of a tobacco tin which he carefully prised out of its connector. When he powered up the machine without the tin box the screen still lit up but the machine behaved differently. The screen went blue and a message appeared to the effect that the machine could not find what it called an 'Operating System'. This told Alan quite a lot. The tin box (which was also the source of the whirring noise that the machine made) was obviously some sort of memory device – probably a miniature version of the magnetic drum storage with which he was already familiar. This gave him confidence that the machine he held on his lap was not qualitatively different from the machine back at the university, only smaller and vastly more powerful. The fact that the screen displayed a message and still responded to the keys on the keyboard showed that it had at least two levels of memory and functionality. But what he really wanted to do was to take the whole thing apart to see what was inside – but this did not appear to be possible without destroying the machine, an option which he was naturally reluctant to choose.

Fortunately, the solution arrived in the form of another parcel which arrived a week later. Inside it was another identical machine but the plastic covers were loose and all the components were visible. There was also another note from the mysterious Bill G. which said 'I figured you would want to see the inside of one of these so have a go at this one. Most of it works anyway.'

Soon Alan was busy with his CRO probing here and there, trying to determine which bits of the machine became operative when the machine was doing different things. But this task proved to be incomparably more difficult than he expected. It seemed as if, when the machine was operating, all of it was equally busy whatever it was doing. He quickly identified the large square object with the cooling fins in the middle of the circuit board as the most important component – the real 'brain' perhaps – because it got hot quite quickly but it was almost impossible to say what all the other black plastic squares were doing. They were obviously connected together with electrical wires beautifully engraved in copper and gold on the circuit board – but what was going on inside them? Mr G had thoughtfully provided a few spare components as well and soon Alan had carefully sliced the top off one of them and had examined the tiny chip of shiny metal which he found inside under a microscope. What he saw astonished him even more. It looked like the plan of New York! There seemed to the streets and avenues, areas of parkland and wasteland too. Then the penny dropped. The streets and avenues were electrical connections and the houses were electrical components – probably miniature transistors, the components that were being used to build the next Manchester computer, the 'Atlas'. One large component seemed to have a particularly regular layout and a quick estimate revealed that it probably contained over a billion transistors, if that is what they were. The sheer complexity of these

devices staggered him and he went back to the invoice to check the date on it. Yes, it really did say 2013, not 3013 or 4013.

“That's only sixty years away. Less than a lifetime. I can't believe that that is possible” he thought to himself; “But it must be possible. Here it is sitting on my desk. It works. I can see that it works. It doesn't use magic. There is no 'ghost' in this machine. It is just logic gates connected together in a fabulously complicated way. If human beings can create a machine as complex as this in sixty years, then perhaps we shall be able to create something as complex as a human brain in another sixty years. Perhaps another lifetime will see the creation of conscious machines. Who knows what might be possible in the future?... ”

The present state of our understanding of the workings of the human brain is almost identical to Turing's understanding of the modern laptop. By cutting bits out of brains and by monitoring the flow of blood within it we know roughly which bits do what and we have some idea how the various areas of the brain are wired up together. Like Turing who was familiar with valves and the newly invented transistors, we know how an individual neuron works, but again like Turing, there is an enormous yawning gap between our understanding of the macroscopic and microscopic workings of the brain. Using the tools at his disposal, I don't see any way in which Turing could figure out the instruction code employed by CPU (though he was familiar with the instruction code used by the Mark I); still less could he determine the syntax of the language in which the laptop's operating system was written in (though the first high-level languages were being developed at the time) Similarly I see little prospect in the foreseeable future of identifying the intermediate levels of organisation that, presumably, lie between say the simple processing that goes on in the visual cortex and the areas of the brain which are responsible for our three-dimensional visual perception of the world around us. Undoubtedly our understanding of the human brain will go on increasing and the gap between our top-down and bottom-up knowledge will get smaller but I have another reason to believe that, with the tools and theoretical knowledge which we currently possess, we will never fully understand the human brain and that is because I believe that the human brain employs processes which are non-classical and that only devices which use these non-classical modes can ever be *conscious*. It is this thesis which I wish to explore.

Theories of the Mind

The essential dichotomy which lies at the bottom of the mind-body problem goes right back to those two founders of philosophy: Plato and Aristotle. Plato is famous for his theory of forms – the idea that the material world, including our own bodies, is but a shadow of a higher realm of ideal forms in which the immortal soul resides. While we live, our soul is, temporarily, attached to our bodies in the same way that the ideal form of a cube is, temporarily, attached to a child's building brick. This idea, in one form or another has enjoyed enduring popularity down the centuries and still forms the basis of most religious philosophies. It provides a satisfying explanation (to

many people at any rate) of the overwhelming sense of self that we all experience whenever we contemplate our own minds as well as conferring other benefits such as providing a moral compass in our relations with other humans (because they, presumably have souls too) and offering the possibility of life after death (because the soul is immortal).

For those of us who do not want to believe in the religious trappings of a soul, it is still reasonable to hold the opinion that science does not yet hold all the answers in respect of how our conscious sense of self arises and that there must exist *something* which will eventually explain the riddle and of which science is either currently only dimly aware (dark energy? quantum decoherence? self-organized complexity?) or, perhaps, completely unaware (????).

It is possible to categorise the various forms of mentalism – which I here take to mean any scientific theory of the conscious mind which emphasises the importance of substances, structures or processes over and above the fairly well established physical processes that go on inside the neurons and synapses inside our brains – in term of what this extra element is.

Cartesian Dualism is the doctrine that the mind is (for want of a better word) a 'spiritual' entity which overlooks the workings of the brain and which can, when required, interfere with it. (Descartes himself thought that it had a definite location in the brain and identified the pineal gland as the seat of consciousness) This 'spiritual' entity was ridiculed by the philosopher Gilbert Ryle in the 1940's who dubbed it the 'ghost in the machine' and few scientists and philosophers would own to subscribing Cartesian Dualism these days because a) there is no actual evidence for it and b) it really does not actually explain anything. Nevertheless, we should not throw it overboard completely. We do need to keep in mind the serious possibility that there may be something in the physical world – a substance, a structure or a process – about which we currently know nothing but which is essential to understanding the mind. Indeed, as will become clear, I myself am of this opinion and therefore could be said to be a Dualist of a sort.

Much more popular these days are theories which view the mind as an 'emergent property' of the brain in the same way that temperature is an emergent property of gas molecules, or nest building is an emergent property of a colony of ants. This position, sometimes referred to as Epiphenomenalism, comes in a wide variety of guises which are distinguished by the varying ways and degrees in which the brain (ie the electrical activity of the neurons) and the mind (ie the mental states which emerge from all this activity) interact with each other. For some, mental states are just a way of describing an extremely complex physical state in the same way that the temperature of a gas is just a way of summarising the average behaviour of a large number of gas molecules. Under this view, mental states are peripheral to the important process in the brain. Mental states are nothing more or less than physical states. For every mental state there is a physical state and vice versa; they are two sides of a coin and they develop in parallel.

This is the position adopted by the electronics engineer who insists that a complete understanding of the workings of a laptop computer can, in principle at any rate, be provided by a complete wiring diagram and a list of all the data held in permanent storage. It is not necessary, the engineer would claim, to know the instruction set of the CPU or what language the operating system is written in; if you know how the machine is wired and the state of its memory, you know, in principle if not in practice, everything you need to know to predict what it will do.

The difficulty with this position when it comes to explaining the workings of a conscious brain is that it is not clear what *use* these 'mental states' are to the organism which possesses it. 'Mental states' may be of use to a psychologist faced with the need to help a mentally ill patient but there does not seem to be any reason why the individual needs to be *conscious* of his mental state any more than a gas needs to be conscious of its temperature.

While parallel epiphenomenalists play down the role of mental states, others (who we might label interactive epiphenomenalists) believe that mental states have an essential role to play in determining what the brain does next. We might, for example, imagine that the action of poking a stick into an ants nest provokes a 'mental state' in the colony called 'stress' which itself causes the ants to behave defensively. It is obvious that a single solitary ant cannot possess this state because 'stress' is a property which belongs to the whole colony. Moreover, unlike temperature which is a simple average over a large number of molecules, 'stress' in an ant colony is more a description of the way in which the colony is organised and in this respect it is more like the concept of entropy than temperature. Now there is a respected body of opinion within the physics community which regards the second law of thermodynamics as a truly fundamental law on a par with Newton's laws of gravity or motion and if this were true, then it would be impossible to explain how a petrol engine or a refrigerator works using the laws of motion alone; you would have to use the concept of entropy. In the same way, the student of ant behaviour would be forced to use the concept of 'stress', not just as a short cut but as an essential tool in understanding and predicting the behaviour of the colony.

Advocates of this position face the same difficulty that confronts the Cartesian Dualist: what is the actual mechanism by which 'mental states' in a brain or an ant colony can influence the behaviour of 'physical states'? If we point to the observation that disturbing the ants by poking a stick into the nest causes some of them to release hormones into the air which triggers a defensive reaction in the other ants, then we have already admitted that the supposed mental state called 'stress' is not the actual cause of the reaction; in which case we can, in principle, dispense with it. Similarly, if we discover that 'being in love' is always associated with exceptional activity in a particular cluster of neurons, then we could, in principle, replace that well-known phrase in the literature with its alternative, but much less poetic description.

Finally, we come to the other extreme position which we might describe as Materialism. Aristotle was a materialist. The (conscious) mind (or, if you like, the

Spiritualism: the belief that human beings possess a completely disembodied immortal soul which is only temporally associated with a particular brain.

Cartesian Dualism: the belief that human beings possess a spiritual entity which is in some way uniquely associated with a human brain and which ceases to exist at death.

Structural Dualism: the belief that the brains of human beings possess unique structures which carry out processes as yet not fully understood by science.

Interactive Epiphenomenalism: the belief that consciousness and its associated mental state is an emergent property of a human brain and that it is impossible to understand the physical behaviour of a brain without taking its mental state into account.

Parallel Epiphenomenalism: the belief that consciousness and its associated mental state is an emergent property of a human brain but that, in principle, the operation of the brain could be understood without reference to it.

Panpsychism: the belief that any artefact capable of any amount of information processing possesses a degree of consciousness.

Radical materialism: the belief that no physical artefact possesses mental states and that consciousness is an illusion.

---Mentalist theories---

---------Materialist theories---------

Fig 1: Theories of the Mind

soul) was simply an attribute of a human being in just the same way that being cuboid was simply an attribute of a brick. The modern materialist will not speak of souls or even attributes; he will go further. He will claim that the whole mind/body dichotomy is a red herring; the mind is not just an attribute of the brain, they are, in fact identical. Mental states are physical states – nothing more, nothing less. If pressed with the objection that my *idea* of a football is not an *actual* football, he might be persuaded to say that the distinction between the mind and the brain is a bit like the difference between the software and the hardware in a computer. If pushed hard enough, the die-hard materialist may be forced into one of two corners. He may admit that, ultimately, he does not believe in mental events at all and, if you are really cruel, you may be able to get him to deny his own consciousness. (When not calling him a fool, I would label such a person as a radical materialist.) On the other hand, he might take the view that, since human computers (brains) possess consciousness, other sentient beings must also possess the same quality but in lesser degree. And if you press this argument vigorously enough, you may get him to admit a degree of consciousness to robots, thermostats and even stones (which feel and respond to the

force of gravity). This point of view has generously been given the grand name panpsychism.

To summarise what we have said so far, many theories of the mind can be pinned on a column which has spiritualism at the top and radical materialism at the bottom (see fig. 1). The attentive reader will notice that I have sneaked in another 'ism in the middle and I may as well admit at the outset that this is where I intend to pin my own manifesto – but first let us start with some incredibly important observations and facts, many of which are conveniently overlooked by adherents of the other 'isms on the column.

5 important observations

1) *Brains are only conscious some of the time*. Brains can be asleep, drugged or in a coma. If we are to understand what makes a brain conscious, we need to study in detail the difference between the conscious and the unconscious brain.

2) *Conscious brains can do things which unconscious brains cannot do*. If this were not the case, there would be no evolutionary need for consciousness. Notwithstanding this argument, we still need to establish exactly what it is that conscious brains can do which other brains cannot.

3) *Conscious beings possess the ability to memorise extremely complex information (including images and sounds) for long periods of time*. I shall argue that, although it would appear to be possible for a creature to be conscious without long-term memory, in practice the former is not of the slightest use without the latter.

4) *Conscious beings report an intense feeling of 'self', of 'being' and of uniqueness*. In other words, consciousness is a *subjective* quality and, on the face of it, this makes it very difficult to reconcile with the *objective* nature of scientific enquiry.

5) *Conscious beings also report a strong belief that they can control the future by carrying out certain actions or not as they will*. The scientific debate on the issue of free will generates such heated responses from both sides that virtually all mind-theorists have completely ignored the subject.

Each of these five observations is telling us something essential about the conscious mind and any theory of the mind which fails to address all these issues is defective in some way. I shall consider each in detail in turn and along the way I shall attempt to shed light on the following central questions to which any theory of the mind should answer:

A) *Are there degrees of consciousness?*

B) What creatures other than human beings possess consciousness?

C) *At what stage in its development does a human child become conscious?*

D) *What are the evolutionary benefits of consciousness?*

E) *Will it ever be possible to attain a proper scientific* explanation *of consciousness?*

F) *Would such an explanation shed any light on the age-old problem of free-will?*

G) *Will it ever be possible to construct a machine which is conscious?*

Observation No 1: Brains are only conscious some of the time.

When we are asleep, we are not conscious. When we are anaesthetized or in a coma we are not conscious. It follows that the mere possession of a brain is not, in itself, sufficient to guarantee consciousness. There are some who maintain that consciousness will always emerge as a natural by-product whenever a system reaches a certain level of complexity but it is clear that complexity on its own is not a sufficient criterion – it all depends on how that complexity is organised.

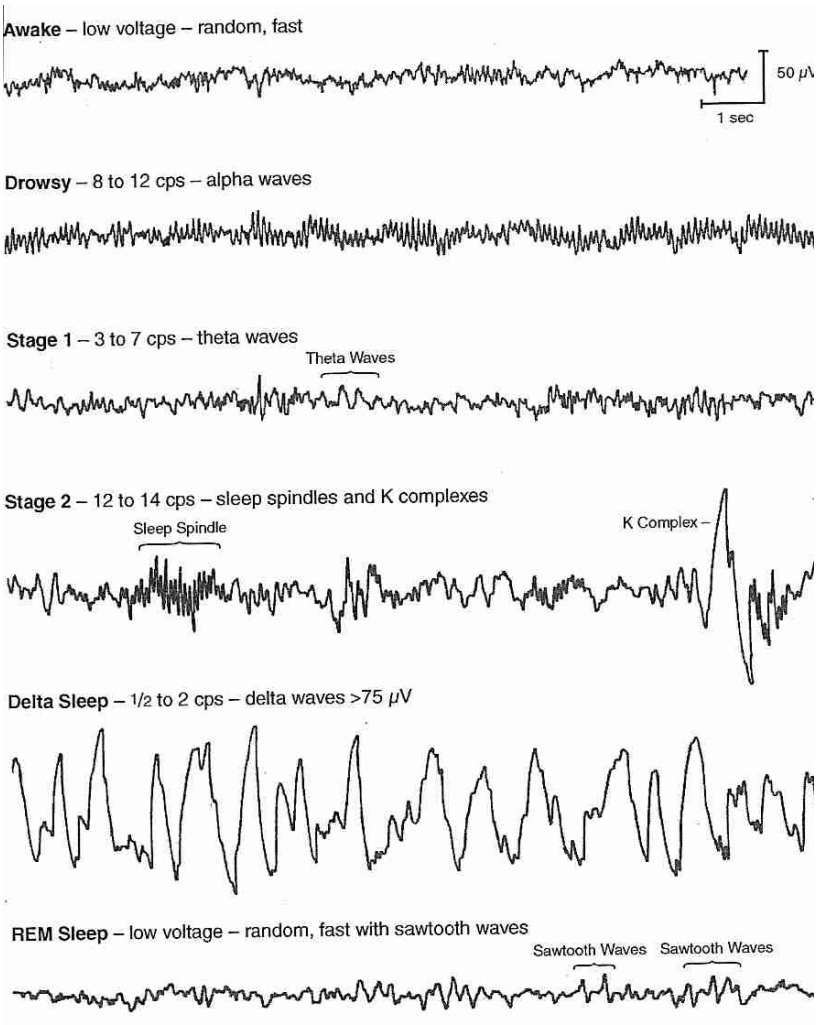

Fig 2: EEG wave patterns (from http://www.benbest.com/health/sleep.html)

Evidence from the electroencephalograph

A crude but important method of observing the activity inside the brain in both the conscious and unconscious states is with an electroencephalograph or EEG and up to 5 different characteristic wave patterns can be observed in the human brain while awake and asleep. (see fig. 2) When the subject is awake and conscious, the EEG records rapid and irregular oscillations of relatively small amplitude. Since the EEG electrodes cover a vast area of cortex (thousands if not millions of neurons) this observation is consistent with the hypothesis that, while the brain is awake, all areas of the brain are more or less active and are 'doing their own thing'.

As the subject begins to fall asleep, the famous 'alpha' rhythm starts to become evident. This is a more coordinated oscillation with a frequency of 8 to 12 Hz. This stage is known as 'drowsy sleep' and the subject may be conscious of his surroundings but unable, or unwilling, to react to them.

The first stage of true sleep is characterised by an even slower oscillation called 'theta waves' of frequency between 3 and 7 Hz. The subject is now truly unconscious (but see below) and over the next 20 minutes or so, descends into two further stages of deep sleep the last of which – delta sleep – is characterised by highly coordinated electrical oscillations of large amplitude and slow frequency.

At intervals during the night, the subject returns to stage 1 sleep and enters what is known as Rapid Eye Movement or REM sleep. The EEG pattern is similar to the waking state with rapid, low voltage oscillations and oxygen consumption by the brain increases dramatically but the subject's muscles are (usually) paralysed and he is more difficult to wake than at other times. As the name suggests, the stage is accompanied by rapid movements of the eye and if the subject is wakened during this phase, he is more likely to report that he was dreaming at that instant.

These, then, are the basic physiological facts. The question which interests me is this. *Are we conscious during REM sleep*? I hear a chorus of replies – but the shouts of "Of course we are!" are almost equalled in volume by those who hold the opposite view. To those of you who responded in the affirmative I ask why it is that, although you entered REM sleep 5 or 6 times last night, you are totally unaware of that fact. Surely if you were conscious during those times, you must have been conscious that you were conscious? And to those of you who replied "of course I wasn't conscious – I was asleep!" I ask you to recall at least one occasion when you had a vivid dream. Were you not conscious of that dream? How can you possibly dream if you are not conscious of what you are dreaming?

We are in danger of tying ourselves in knots here but the point is a really important one. All the physiological signs indicate that we are indeed as conscious during REM sleep as we are when we are awake; the difference being that certain functions (like the ability to move our muscles at will and to respond to stimuli) are deliberately suppressed (probably to save ourselves from self-harm). If, at the same time, our memory circuits also suppressed, that would explain why we are so rarely able to

remember our dreams and why we cannot recall being conscious during the night.

It would seem to me that, in this instance, we must accept the physiological evidence from the EEG machine over our subjective experience. During REM sleep the brain seems to be working overtime. We don't know what it is doing but it certainly appears to be doing the same sort of things that it does when we are awake. I know that I was conscious yesterday evening because I remember watching the 10 o'clock news. The fact that I cannot remember what I was thinking 2 hours later is not proof that I was not conscious then – only that I cannot remember the conscious thoughts which I had at that time.

If, therefore, we accept that we *can* be conscious during REM sleep (but never during non-REM sleep), it would seem to be highly likely that other animals which also show similar patterns of EEG activity during sleep are also conscious.

Many, if not all, animals sleep – including invertebrate species – but nobody is quite clear why sleep is necessary; indeed, it would appear to be a rather risky option. Sleep is not a problem for predators at the top of the food chain or animals which can hide themselves effectively but it can put other animals in serious danger. Some animals (eg monkeys) live in social groups so that some members can sleep while others keep watch; others live in large herds for much the same reason. Marine mammals such as dolphins and whales have come up with another solution: they sleep with only half of their brains at one time! Seals can do both. They can sleep with one half of their brain while out at sea, but sleep with both halves while safe on land. Many birds also employ unihemispheric sleep and it is said that migratory birds can sleep on the wing. This has not been conclusively proved for obvious reasons but I see no reason to doubt it because it is only the areas of the brain associated with consciousness which shut down during deep sleep. If sleepwalkers can get up and make a cup of tea without being conscious of so doing, I see no reason why a swallow cannot fly in its sleep. (Sleepwalkers are not 'acting out their dreams' as used to be thought; their EEG patterns are those of non-REM sleep, not REM sleep.).

Now it is a fascinating fact that while most animals sleep, only mammals and birds exhibit REM / non-REM cycles of sleep in greater of lesser degree.

Reptiles need sleep as well as birds and mammals but their EEG waves do not show any evidence of a REM like phase. This seems to suggest that sleep in reptiles is more a way of passing the time and giving the body a rest than anything to do with the demands of the brain. Fish too sleep, but their brain activity is difficult to record.

Some form of sleep appears to be necessary even for insects and crustaceans but although depriving these creatures of sleep impairs their ability to learn, measurements on their nervous systems during sleep shows no evidence of a REM type phase.

If, then, we go along with the idea that REM sleep is indicative of consciousness, then all mammals and birds are conscious in some degree but reptiles are not. If this

is true, it raises an interesting question with regard to the evolutionary development of these families. The common ancestors of these groups are small lizard-like creatures called amniotes which lived in the late carboniferous period some 300 million years ago. Their eggs were encased in a sack containing amniotic fluid and this enabled them to reproduce on dry land without having to return to water. This evolutionary branch quickly divided into two, the synapsids (which developed into mammals) and the sauropsids (which became reptiles, dinosaurs and birds). Now since, according to my thesis, reptiles are not conscious, this would seem to imply that consciousness has evolved separately in mammals and birds. It would also appear that there is a strong correlation between blood temperature and consciousness. Indeed, judging by the fact that the human brain uses 10 times as much oxygen per kilogram as the rest of the body, I would go so far as to suggest that being warm-bloodied is a necessary condition for consciousness. (The fascinating question as to whether any of the dinosaurs were conscious will probably turn on whether or not they were warm-bloodied.)

Evidence from anaesthetics

Before we leave this highly instructive topic, is there anything to be learnt about consciousness through studies of anaesthetics? When we go into hospital for a major operation under general anaesthetic, we do so under the expectation that, however much the surgeon cuts, slices and stitches up our bodies, we will feel no pain at the time and will emerge from the theatre with no memories of the experience whatsoever. We explain this to ourselves by saying that, during the operation we are simply unconscious. But how do we know this? I have argued that, during REM sleep (and specifically while dreaming) the brain is in a conscious state – but we are unaware of the fact because when we wake up in the morning we usually have no memories of the experience. Could it be the same during open heart surgery? Do we, in fact, feel every incision of the knife, every snip of the scissors, every prick of the needle at the time, but simply have no memories of the ghastly experience when we wake up? What a terrifying thought! How do we know that we are unconscious during non-REM sleep? How do we know that the patient in a coma is unconscious?

The short answer is that we don't. But we must not ignore what little objective evidence there is. EEG studies show us that the brain can exist in one of several recognisably different states of activity. One of those states (the rapid, random electrical oscillations associated with being awake or in REM sleep) is definitely connected with the subjective experience of consciousness. The other states are associated with periods of which the subjects later report having had no conscious experience. Why should we doubt them?

Consciousness and pain

There is another way in which we can judge whether a person is or is not conscious

and that is by studying their response to painful stimuli. Of course, even when asleep, the brain is constantly monitoring its surroundings and carrying out primitive remedial actions in response to stimuli. If you shine a light on a sleeping person, they will probably turn over and bury their head under the pillow; if you remove the bedclothes they will probably curl up to keep warm; if you make an unusual noise like the sound of breaking glass, they will probably wake up. None of these responses requires action from the conscious parts of the brain. Even if you inflict pain, for example by pricking them with a needle, the sound sleeper will probably react by merely withdrawing the limb. What they will not do is sit up and say “Ouch! that hurt! What did you do that for?”. (Even if the subject is enjoying REM sleep at the time and who is therefore, by my theory, conscious will probably not sit up and complain either because, for some reason, subjects in REM sleep are even more difficult to wake up than subjects in deep sleep. The difference comes later when you ask them what happened during the night. The deep sleeper will have no recollection of the event at all but the REM sleeper will say “It's funny you should ask about that. I had this curious dream in which I was in a jousting tournament and I got stabbed in the arm ...”)

It is now accepted that pain has evolved because it has survival value. If you accidentally put your hand on a hot surface, the pain you experience will rapidly cause you to take appropriate action to withdraw the hand from the source of heat. Notice that this is not the same as the familiar knee-jerk reflex which is not under the control of the brain; this requires a response from much higher up the nervous system. In fact, it would appear that pain goes, as it were, right to the very top and that, in order for it to be of any use as a survival mechanism, the subject has to be conscious in order to experience and therefore to react to pain. It follows therefore that, with the sole exception of subjects in REM sleep, if the subject fails to produce any of the usual responses to painful stimuli that a conscious person would produce, the subject must be unconscious. Sleepwalkers are pretty oblivious to pain and can do themselves serious harm. We are therefore right to conclude that they are unconscious – a conclusion supported by evidence from their EEG patterns.

If we apply the same test to animals, it is immediately apparent that all mammals show exactly the same difference in response to painful stimuli when they are awake and when they are asleep as humans do. If you tread on a cat or kick a dog, it complains. So do rodents and herbivores. Cats and dogs, rats and cows can be anaesthetised using exactly the same drugs as are used on humans. There is little room for doubt. All mammals can experience pain and therefore all mammals are conscious (some of the time and in some degree).

Although the evidence is more difficult to obtain, birds too can be anaesthetised but it is less clear how their responses to painful stimuli change under these circumstances. Nevertheless, I think it would be unwise to assume that, just because we do not understand their body language as well as we understand the body language of other mammals, birds are not capable of suffering pain.

I imagine that fish can also be 'put to sleep' using drugs but I doubt whether it is necessary to go any further than appealing to the analgesic (pain-numbing) rather than the anaesthetic (sleep-inducing) properties of the drug to explain any changes in the behaviour of the fish.

Other animals such as insects, molluscs and crustaceans show little evidence that their response to traumatic stimuli can be changed reversibly by anaesthetic drugs so I think we can be reasonably confident that they do not experience pain and are therefore not conscious in any sense which implies a degree of similarity with what humans describe as consciousness (though there may be an exception to this rule in the case of cephalopods such as the octopus).

If my analysis of which animals are conscious and which are not is correct, the ethical implications could be profound.

Consciousness in human children

It is time now to raise the extremely emotive question of when, in its development from embryo to child, does a human being become conscious. First a few facts: the brain starts to develop after about 8 weeks gestation; by 22 weeks, the foetus shows certain primitive reflex actions and a few weeks later the central nervous system is fairly well developed. By 32 weeks the foetus can probably see and hear, smell and touch (though obviously *what* it can see etc. is rather limited!). At birth, the infant brain is fully developed in so far as all the relevant structures are there, its EEG patterns are much the same as for an adult and its responses to painful stimuli can be modified by anaesthetics – but is it *conscious*?

There can be little doubt that the answer is yes and that the faculty of consciousness develops rapidly during the last 10 weeks of gestation. However, lacking any experience of world outside the womb, it cannot be conscious of very much and whether or not it is in any sense conscious of *itself* is a question to which I shall return.

Observation No 2: Conscious brains can do things which unconscious brains cannot do.

I think that most of us (the panpsychist excepted) would agree that computers are not conscious. But the list of things which computers and computer-controlled machines can do is impressive and likely to become even more so. Computers can beat almost anyone at chess; they can diagnose illnesses; prove mathematical theorems; build cars; guide missiles to a target; explore distant planets etc. etc. etc. But no computer has (yet) invented a new joke, written a decent poem or composed a symphony. These examples seem to suggest that what a computer lacks is the ability to *imagine* and *create* new things which have never been imagined or created before. Indeed, if I were asked to adjudicate in a Turing test between a computer and a human being, that is what I would ask the computer/human to do – create something. Of course, there

are many human beings who would not pass this test (myself included) but that is not the point. If, as a result of my request, the terminal printed out a really novel joke or the score of a brand new symphony, I would conclude that the being behind the screen was human.

Having said that, I do not entirely rule out the possibility that a computer made of wires and silicon could never pass this test or that a machine which passed this test was necessarily conscious. All I am saying is that, with our current level of technology, there are still certain things which conscious human beings can do which computers cannot.

What then of supposedly unconscious animals like insects and crustaceans? Are there things which mammals and birds can do which these creatures cannot? Do mammals and birds show any evidence of imagination or creativity? Are there problems which insects just cannot solve? What about fish? Do they fit into the pattern?

I am no expert on these matters but on the whole I think the answer is yes. One of my favourite stories (which is probably apocryphal) is that of the chimpanzee who was faced with the problem of retrieving an apple floating just out of reach in a bucket half filled with water. The experimenters wanted to see if the chimp would realize that, by dropping a brick in the water, the water level would rise sufficiently to allow him to reach the apple. They were astonished when the chimp ignored the proffered brick and instead reached the apple by peeing into the bucket! Ravens can solve similar problems but can you imagine a spider or even an octopus (supposedly the most intelligent cephalopod) doing this? Douglas Hofstadter describes the behaviour of the Sphex wasp which repeatedly checks its nest over and over again whenever the experimenter disturbs its routine by moving its catch away from the nest. It is clear that the wasp does not have the *imagination* to modify its routine to cope with the new situation.

Now it might be argued that a spider, faced with a garden shed, a spade and the branch of an overhanging tree, wishing to build a web, is faced with a serious problem which requires imagination and creativity in its solution – but you would be wrong. It would be a relatively trivial matter to program a computer-controlled robot to do this task.

So what kinds of problems do mammals and birds have to solve that a computer would find really difficult? This is not an easy question to answer but since, presumably, consciousness has evolved because it gives its owners a competitive edge, the answer must involve either finding food, finding a mate or rearing young. Now many mammals (including whales) and birds travel vast distances in search of abundant food or good nesting sites, so having a conscious brain may be of use to them in performing the necessary navigation. It is certainly difficult to explain how a pigeon can find its way home even when it is transported hundreds of miles away from its roost; elephants can return to a water hole after many years of absence and many birds return year after year to the same nesting sites. How do they do this? It is

tempting to suggest that they have a conscious awareness of their surroundings.

On the other hand, monarch butterflies migrate thousands of miles, cruise missiles do exactly what homing pigeons do and even the humble limpet returns to the same spot on its rock after a days foraging so this is no proof that mammals and birds are conscious. And in any case, this sort of feat does not seem to require imagination or creativity.

Perhaps imagination is required when an animal is required to change its normal habits as a result of habitat loss or climate change. I don't think anyone will argue with the thesis that the success of mankind in dominating this planet is entirely due to his unique ability to adapt his behaviour according to the circumstances he finds himself in and that this adaptability springs from his ability to use conscious reasoning but it would be difficult to extend this argument to other putatively conscious animals.

I think the real answer must lie elsewhere. The one feature that really distinguishes mammals and birds from fish, crustaceans and insects is their ability to *recognise each other as individuals*. This is obviously true of the apes and I believe it to be true of elephants, whales and dolphins too. All social animals such as lions and wolves need this skill as do all birds who mate for life. The hypothesis is more difficult to prove in the case of smaller mammals and birds but any creature who suckles its young or feeds a chick would be well advised to be able to recognise its own offspring. (I am afraid that the poor willow warbler who feeds a cuckoo chick twice its size cannot be credited with much imagination and it is in all likelihood not conscious of what it is doing!) I think it extremely unlikely that any insects or crustaceans have the ability to recognise other members of the species as individuals and I would be very interested to know if any fish (such as sharks, which definitely hunt in packs like wolves) can do it. My gut feeling is that sharks, like bees or ants, behave cooperatively in the hunt by instinct and whereas I can imagine a wolf thinking to itself (in wolvish) 'loppy-tail is a fast runner, I will leave the chase to him and go round the back here to cut off the deer's retreat', I can't see a shark thinking like that.

My conclusion is that the possession of a conscious mind permits creatures to have *personal relationships* and it is this, more than anything else which gives mammals and birds their evolutionary advantage.

Observation No 3: Conscious beings possess the ability to memorise extremely complex information (including images and sounds) for long periods of time.

It is almost as difficult to understand how our memory works as it is to understand consciousness. Indeed, I suspect that if we knew the answer to the former question, we would be well on the way to answering the latter. It is pretty well established that the human brain does not store memories in the same way that a computer stores

information. There is no single collection of neurons in your brain that holds your credit card PIN number. The metaphor of a hologram is probably more helpful or even that of a fractal algorithm which somehow enables you to reconstruct an image of a fern. It should also be remembered that humans have several different kinds of memory and it is probable that different methods are used to store information in each case. Short-term memory – the memory that you use to write down a telephone number a few minutes after you have been told it and the kind of memory that I always appear to use whenever I am told the name of a new acquaintance! – is probably dynamic in the sense that it requires the continuous firing of certain neurons and is almost instantly forgotten. Long term memories, on the other hand, are probably held as a result of almost permanent changes to the way that the neurons in your brain are connected together.

Now it is often said that 'elephants never forget'. I don't suppose that elephants are really any less likely to forget things as we humans are but what is indubitable is that they can remember things for a long time. I have already mentioned their ability to remember the location of a water hole last visited many years ago and the sight of a young elephant trumpeting over the bones of his mother killed by poachers months before is poignant testimony to their ability to remember past events. Dog handlers will recount stories of impressive feats of memory by their pets and penguins can recognise their mates after months of separation at sea. Recent research suggests that dolphins can remember the calls of individuals which they last met as long as two decades ago. Even rats (who have become familiar with several different mazes) can remember where the bait was placed last time they ran the maze for at least a week.

Nothing here suggests any necessary link with consciousness, though, and it may indeed be the case that the two phenomena are either entirely distinct or more probably two products of the same feature of the brain but it does seem at least plausible that consciousness and long-term memory are closely linked and that you cannot have one without the other. If, as I have suggested above, the evolutionary advantage of consciousness is to enable the creature to engage in long-term personal relationships with other members of the species then there is no point in being conscious if you can't remember what you were once conscious of.

I would be interested to know of any evidence that any of the lower orders of animal possess a long-term memory. Apparently French angelfish form lifelong bonds but this is very rare and can probably be explained without any reference to conscious memory but simply in terms of adaptive behaviour. I would be extremely surprised if, after a period of separation, two angelfish would necessarily resume their former cooperative relationship with each other.

Towards a definition of consciousness

So far we have discussed three ways in which we can objectively identify the probable occurrence of consciousness in creatures other than ourselves; but we have

not begun to address the question of what consciousness actually *is* or how the subjective experience of being conscious actually comes about. In fact, I believe that we are a good deal further from understanding the mechanisms of consciousness than my fictional Turing was from understanding the workings of a computer. In respect of our understanding of consciousness, we are more like Babbage than Turing. At least the latter grasped the fundamental principle that his laptop used electricity; Babbage would have been puzzled to find a complete absence of gear wheels and cogs inside the machine!

Are there any pointers at all to where we should start looking for the extra ingredient which is necessary to convert an unconscious brain into a conscious one? Is it just a question of organization (as the epiphenomenalists would have it)? Or are we going to have to take paranormal phenomena like mind-reading and telekinesis seriously in our search for the key to consciousness (as the spiritualists would have us believe)?

No open-minded scientist should rule out either possibility, but there is another alternative and that is that there is some perfectly rational physical process going on in the conscious brain of which we currently have no understanding whatsoever. Lightning was considered an act of God until Franklin showed that it was just a form of electricity; Heat was considered to be a type of fluid until Joule showed than it was a form of energy; Gravity was thought to be something called a 'force' which propagated instantaneously over vast distances until Einstein showed that it was due to the curvature of space-time. And if astronomers can trump all the evidence patiently collected by our particle physicists using their vast and expensive accelerators by proposing that most of the universe is composed of something which nobody has ever seen just on the basis that a few incredibly distant galaxies are a bit dimmer than they ought to be, surely I can be forgiven for proposing that, since the phenomenon of consciousness cannot be adequately explained by existing scientific theories, it must be because it relies on some property of nature which we have yet to discover.

Now it is tempting to suggest that, since there are in fact two things which we fundamentally do not understand – namely consciousness and quantum theory – the solution to one riddle may lie in the other. In fact, it has been said that this is the only argument in favour of a quantum theory of the mind but I have to disagree. I believe that there is at least circumstantial evidence to support this view. When two particles enter an entangled state they could be said to have, temporarily, become one entity with properties quite different from the two particles which they eventually turn back into. Perhaps bits of our brains have the capacity to enter quantum states with similar non-classical properties. Perhaps our conscious brains are quantum, not classical computers and it is this which gives them the power to dream up new jokes, invent new mathematics and compose symphonies. Although this idea may seem far-fetched, it would, at least, explain how the same brain can sometimes be conscious and sometimes unconscious. When it is conscious it is operating in quantum mode; when unconscious it is operating in classical mode.

One minor objection to this idea is that the process of creativity (which I have assumed is the hall-mark of a conscious creature) often seems to take place in our unconscious minds. I am told that Mozart would wake up one morning with a complete symphony in his head; and we all know that when we have a particularly knotty problem to solve, the best thing is to go and do something else and often the answer will pop into our heads by itself. I do not regard this objection as serious. I have not claimed that conscious experiences *necessarily* accompany quantum processes in the brain and I deem it perfectly possible that the latter can occur without the former. It seems entirely likely that there are degrees of consciousness and that, even while we sleep, there may be parts of the brain which continue to operate in quantum mode (e.g. during the REM phases). Since, as I have argued above, consciousness also appears to be intimately associated with long-term memory, I would also like to suggest that the subjective experience of consciousness is not just a product of quantum processes in the brain but also necessarily involves the presence of sensory inputs which are interpreted by the conscious brain in terms of its past experiences. The implications of this are quite profound. I am suggesting that the new-born infant, although in possession of a fully functioning quantum brain, can only really be said to be conscious in a technical sense. Since it has no memories of past experiences with which it can interpret the sudden change in sights, sounds and smells which suddenly assault its senses, it cannot truly be said that it is conscious *of* them. To put it another way, consciousness without past experiences is not true consciousness at all.

Putting all of these ideas together I would like to suggest the following tentative hypothesis: *consciousness is a result of a particular way in which a brain uses non-classical operations to process incoming sensory information in the light of past experiences to generate novel solutions to problems posed by its environment*.

This places the study of consciousness on a firm scientific basis. If we eventually do discover a structure or process in the human brain which is uniquely correlated with conscious experiences (even if it turns out to be a perfectly classical one) we will have learned a great deal about what causes consciousness and we should be able to use this knowledge to predict the behaviour of other creatures which do, or do not possess the same structures.

But even if we were to reach this stage of enlightenment, would we be in a position to *explain* our *subjective experience* of consciousness, our sense of *being*, our sense of our own *uniqueness*? I am afraid that the honest answer to this is no. But then it is not the job of a scientific theory to explain *everything.* Newton's law of gravity can be used to explain why planets move in ellipses and it can be used to predict the motions of the planets with uncanny accuracy; but it says precisely nothing about why massive bodies attract each other in the first place.

Even if we knew so much about the unique structures and processes that go on in a conscious brain that we could construct conscious machines out of string and sealing-wax, we still would not know *why* our machine was conscious. It follows that we

human beings will *never* understand the workings of our conscious minds in any scientific sense. The best we can hope for is that an increased understanding of the physical laws which govern matter will give us an insight into the way in which the conscious brain interprets sensory information in the light of remembered experiences.

So the only scientific question which is worth asking at this stage is the following: *what are the unique structures and/or processes in the human brain which are uniquely correlated with consciousness and do these structures or processes really carry out non-classical (e.g. quantum) computations or not?* In other words, is the conscious brain a non-classical (e.g. quantum) computer or is it merely a very complex classical one?

The great majority of neurophysiologists work on the assumption that it is the latter because all the micro-structures and processes in the brain which have been observed so far can all be explained adequately in classical terms. But there are a few intrepid scientists who are prepared to think the unthinkable, the most prominent of whom is Sir Roger Penrose who has made a powerful case for the non-classical nature of the brain and who has even identified some parts of the brain where these non-classical processes might occur. In his book 'The Emperor's New Mind' Penrose uses Gödel's theorem to 'prove' that the human brain cannot be a classical computer. Now it must be conceded that many if not most scientists do not accept that Gödel's theorem has anything to do with the human brain but, even if his logic can be questioned, his conclusion may still be correct. Conscious human brains do appear to be able to perform feats which it is extremely difficult to imagine computers performing *however powerful they may become* in the future. One of these feats is the ability to analyse itself by being *self-aware*.

Observation No 4: Conscious beings report an intense feeling of 'self', of 'being' and of uniqueness.

When philosophers and scientists talk about consciousness, one word crops up more than any other and it is: *awareness*. But if any single word has generated more misunderstanding and confusion in the subject than this one, I know it not. It is high time that we abandoned the use of this word and adopted a scale of (semi-technical) terms to describe the various levels of awareness that different systems, both organic and inorganic, display. Might I suggest the following list?

> *Susceptibility*: I use this term is the basic sense to describe anything that is susceptible to an external influence. For example, the Moon could be said to be aware of the Sun because it is *susceptible* to the Sun's gravity. A badly built house could be said to be aware of earthquakes because it is *susceptible* to being shaken.
>
> *Irritability*: this term is used by biologists to describe the awareness that plants and primitive animals show when they respond to changes in their

environment. For example: plants may open their flowers when a light is shone on them; woodlice retreat into the darkness; bees swarm when their hive is disturbed. I think we can also extend the use of the term to describe the behaviour of any man-made device which has sensors and which uses the information from these to control its actions. Such machines would range from a simple thermostat to Google's driverless car

Sentience: this is defined as having the power of *conscious* perception through the senses. Just look around you and be aware of your surroundings. That is *sentience*.

Self-awareness: Look at your hands; touch your face; recognise that in an important sense, these things are different from the other objects around you; they belong to *you*. That is *self-awareness*.

Auto-consciousness: close your eyes and try to think of your own consciousness. If you can do that, you can be said to be aware of your own consciousness i.e. you are *auto-conscious*.

It is probably clear from what I have said earlier that I regard insects, crustaceans and other primitive animals with rudimentary brains as irritable, not sentient (but concede, this is far from proven).

When it comes to categorizing conscious beings, the situation is more difficult. There is good evidence that primates, elephants and grey parrots are capable of self-awareness (paint a spot on their forehead while they are asleep and then put them in front of a mirror.) but other creatures may also be self-aware without passing this particular test. At what point does a human infant become self-aware? That is an important and interesting question.

Another interesting question is this: are all conscious beings auto-conscious? Or to put it another way, is it possible to be sentient or even self-aware without being aware of your own consciousness? Just because we humans find this difficult, it does not mean that it is logically impossible. It is fashionable these days to play down the difference between human beings and the higher primates to the extent that the latter are now afforded certain legal rights in some countries. But if it could be shown that, while higher primates are self-aware, only humans are auto-conscious, would this change matters?

But there is another dimension to conscious awareness which I have not mentioned yet and that is our awareness of the passage of time and of our own immediate past. I think we should call this t*emporal-awareness*. In some ways, I think this is the most important aspect of our conscious awareness and is the reason why I included the phrase 'in the light of past experiences' in my definition of consciousness. To my mind, a new-born baby cannot be said to be truly conscious because its past experiences are so limited. To be fully conscious you have to be sentient, self-aware, auto-conscious *and* temporally-aware.

But this neat classification of degrees of awareness has its problems. In what state of awareness is the dreamer or the hallucinating drug addict? EEG test appear to show that both are conscious in the sense that their brains seem to be doing the same sort of things that brains do when they are awake – but they cannot be said to be sentient let alone self-aware or temporally-aware. The brain appears to be doing its quantum thing without reference to either current sensory data or past memories. We could usefully describe this paradoxical state as being *conscious but inconscient*.

This is all very well but inventing pretty definitions does not prove anything. I agree. But it does help to sort out what is important from what is not. The big question is – does our ability to be sentient, self-aware, auto-conscious and/or temporally-aware shed any light on the processes that are going on in our brains? Would it be possible to design a machine to be any of these things? The epiphenomenalist would argue that all these things are possible if the system in question is complex enough and designed in the right way. My feeling is that there is something qualitatively different between sentience and irritability which cannot be bridged by a mere increase in complexity or subtlety of programming. But in truth I cannot refute his position convincingly. I just don't see how a classical computer could be aware of itself.

But I can hear the epiphenomenalist's triumphant reply: 'I don't see how a quantum computer could be aware of itself either!'

So lets look at this objection more closely. Is there anything in quantum mechanics which could lead to self-awareness or any of those other types of awareness associated with consciousness? I believe there is.

One of the most important aspects of quantum theory is its essential non-locality. In classical physics, all causes are local. The moon moves in such a way because it is responding to local changes in the curvature of space-time; a photographic plate records and image because it is hit by an electromagnetic wave of energy; a billiard ball changes its direction of motion because it is struck by another ball etc. etc.. But in the quantum world things seem to happen either without a cause (eg the decay of a radioactive atom) or because a measurement is made somewhere else (as in experiments on entangled particles). Now I do not wish to get involved in all the arguments concerning the interpretation of quantum mechanics here but whatever interpretation you adopt, you have to conclude that it is not possible to treat any quantum system in isolation; it is always fundamentally connected to its environment. If we adopt the hypothesis that the conscious brain operates on quantum not classical principles, then we would expect that conscious processes in the brain would also be essentially non-local. In short, we might expect conscious thoughts to take place in the whole brain rather than separately in different bits of it.

Now it is a curious fact about our conscious brains that we can only think of one thought at a time. Consider what happens when you are walking along a road, deep in conversation with a friend. The time comes when you have to cross the road. The conversation stops. You assess the traffic and cross when the road is clear. The

conversation resumes. Why did you not continue the conversation while assessing the traffic? The answer is that both processes required the whole brain to compute. It just is not possible to think two different thoughts at once. On the other hand, a computer controlled robot would have no difficulty at all in carrying out both processes simultaneously (It could do this either by employing two microprocessors independently or by time-sharing, it doesn't matter which.) but this is not, in general, possible for a human. (It is conceivable that patients with a surgically 'split' brain could think two different thoughts at once and psychological studies of such patients could shed much light on the potential ability of one half of the brain to communicate with the other by non-local quantum processes.)

Now it is this single-mindedness of the human brain which is responsible, I believe, for the overwhelming sense that we are each a unique individual and I find it highly suggestive that this single-mindedness is a necessary consequence of the hypothesis that the brain is a quantum computer.

And there is another reason why it is extremely tempting to believe that the human brain is a quantum not a classical computer and that is to do with the existence or otherwise of *free will*.

Observation No 5: Conscious beings report a strong belief that they can control the future by carrying out certain actions or not as they will.

The question of whether human (or other) beings possess *free will* has been endlessly debated over the centuries and I am not going to attempt a summary of all the arguments here. One thing is clear, however: your position on this questions is closely related to the issue of where you stand on the mind/body question. If you are a materialist or an epiphenomenalist, our current understanding of the physical laws would seem to indicate that the behaviour of a human brain must either be entirely deterministic (if it is a classical computer) or at best, partially random (if it is non-classical). In neither case is there any room for any process which we could reasonably call *free will*. If, on the other hand, you are a Cartesian Dualist or a Spiritualist, there is no problem. The 'ghost in the machine' can take over whenever it likes and override the laws of physics. As a rational scientist, I am extremely reluctant to adopt this position – but I am equally reluctant to abandon my powerfully felt sense that I can control my future. Is there a middle way?

A glance at figure 1 will reveal another 'ism pinned to the column which I have called Structural Dualism. This is the belief that the brains of human beings (and other conscious creatures) possess unique structures which carry out processes as yet not fully understood by science – the implication being that it is these processes which are responsible for our conscious awareness and also of our belief in our ability to determine the future to some degree.

From what I have said in the previous section, no scientific theory, whether classical or non-classical will, in my opinion. ever explain the subjective *experience* of

consciousness; but some interpretations of Quantum Theory do open the door to at least the possibility of explaining how a conscious brain can exercise *free will*. It might work like this:

We already know that systems of particles can enter what are called *superposed* states. These are situations in which the system can appear to be in more than one classical state at once. Consider a very simple system in which a single photon impinges on a half-silvered mirror. The photon has a choice. Either it passes through the mirror or it is reflected off it. In a classical world, the choice would be made at random and an instant after would be found *either* transmitted *or* reflected. But in a quantum world, this is not the case. Countless experiments have proved without doubt that the photon enters a superposed state in which it is *both* transmitted *and* reflected! It is true that, eventually, when the photon is actually detected by a photographic plate or a CCD device, it turns up in only one of the two places but, for a while at any rate, it really does seem to be in two places at once. Erwin Schrödinger, one of the founders of modern Quantum Theory tried to ridicule this idea by suggesting that, if this were true, a cat could be both dead and alive at the same time but his bluff was called and this is exactly what some interpretations of Quantum Theory affirm.

A more attractive possibility is that when a photon (or a cat) enters a superposed state, its state is temporarily *undecided*. It has the possibility of being in either state but Nature has not yet worked out which state it is going to be. This becomes apparent sometime later when the particle (or object) has interacted with a sufficiently large amount of its environment (a process known as *decoherence* or *objective reduction*). What causes this collapse (if, indeed, it actually occurs) is unknown. It might be something to do with the quantum nature of gravity (as Sir Roger Penrose believes) but I prefer to think, along with many others, that it is brought about by the ever increasing complexity of its relation with the environment with which the system inevitably comes into contact. At the risk of sounding a bit absurd, you could even talk in terms of the environment exercising its *free will* to make a (conscious!) *decision* about whether the photon should be transmitted or reflected.

Now while I would not go so far as to claim that photographic plates have free will, the idea does raise the intriguing suggestion that some exquisitely organised structures the size of a human brain could actually enter into an superposed state in which the whole structure is employed in some sort of quantum calculation. If the outcome of that calculation includes the possibility that a cup of tea is made, and also the possibility that a cup of coffee is made, and if the quantum processes which are going on in my brain are at the same time somehow responsible for my feelings that '*I* ' am making a conscious '*decision*' then surely we are justified in saying that when the wavefunction collapses and I end up making a cup of tea, the '*I* ' in my brain has exercised its '*free will*'.

The usual objection to an argument of this sort is that, if the decision is made by any

sort of natural, physical process, whether it is quantum or classical, the result must be either determined in advance or completely random. In neither case is it possible to claim that my brain has exercised its free will in any meaningful sense.

My answer to this is as follows: Nobody could predict the outcome in advance *even in principle* (because the process is a quantum one) so the process is not deterministic; but neither is it random because the whole, conscious brain (which is the '*I* ' bit) was definitely responsible for causing the one outcome and not the other. The situation can be likened to a General Election. The outcome is unpredictable, but it is not random either because the whole electorate is involved in making the decision.

Some will object that I am just mincing words here but I do not believe that that is so. If conscious creatures possess this thing called free will and unconscious creatures do not, then it ought to be possible to detect this difference in their respective behaviour. To some extent we have already discussed this with regard to the evolutionary benefits which consciousness confers but is there any way we can detect whether or not a creature has *free will*? With humans, it is not a problem. When my wife says 'I am going to make a cup of tea.' and then subsequently makes a cup of tea, we can be reasonably confident that she has made a conscious decision to carry out that action. (Of course, a computer-controlled robot could be programmed to do the same thing in in specific circumstances but in an open ended Turing-style test, its limitations would soon become apparent.)

With creatures that cannot speak, the task is more difficult but we are still looking for the same thing – evidence that the creature knows what it wants to do before it does it. The dog that drops a ball at its owner's feet clearly knows what it wants to do and his posture speaks of his intentions as loudly as words. The squirrel that hides a nut in the forest does so because it knows that it may need to return to the same place in the future. The Jay that watches the squirrel hide the nut does so because it intends to dig up the nut as soon as the squirrel is gone. The homing pigeon that orients itself with respect to the sun and heads off in a certain direction clearly knows that food and safety await it if it chooses the right direction.

But does the spider know what it is about to do when it lays down the first strand of a web? I doubt it because there is no need for it to know. Web-building is an instinctive pre-programmed task. The rules are simple. Place a leg here, place a leg there, draw a thread from A to B, move along one, repeat. When a pike disguises itself among some rocks, is it thinking to itself 'I will just hide in here until a minnow comes along so I can eat it' I don't think so. Its behaviour is entirely instinctive (or, perhaps, learned). Pikes which do not adopt this strategy do not survive.

So why is it that the same creatures which show REM / non-REM cycles of sleep and which can display creative solutions to problems and which possess the ability to remember for long periods of time are exactly the same creatures which show evidence of intent? The answer is that it is because their brains have the ability to use

quantum (or other unknown) processes to remember information and make creative decisions about what to do next. In other words, these creatures have the capacity of conscious thought.

The unconscious brain

If explaining consciousness is (currently) beyond our reach, what about explaining how unconscious minds work? How do ants walk? How do dragonflies hunt? How do snails find food? How do crabs fight? How do spiders make webs? How does a sleepwalking human make a cup of tea?

Well the first thing to say is that we can build robots to do all of these things so the temptation is to conclude that all these creatures are merely complicated robots, pre-programmed by inheritance or learning to do certain tasks. But is this really true? Are unconscious creatures built on the same principles as a modern laptop? Are we in the same position with respect to these brains as Alan Turing was when he looked inside the machine? I would like to suggest that in some respects our task is probably easier than Turing's.

There are two ways to build a robot. You can make a simple white line following machine with nothing more than a couple of optical sensors cross-wired up to two electric motors in such a way that if the sensors drift off to the right, the motors cause the robot to turn left and so on. The circuit is pretty simple and an example is shown in fig. 3.

The second way is to replace the simple amplifier (labelled L293D in the figure) with a programmable chip like a PIC. This can be programmed to do exactly the same job but it has the advantage that it can be reprogrammed to different things like following a black line instead of a white one etc.. The first machine is entirely hard-wired while the second relies on a stored program.

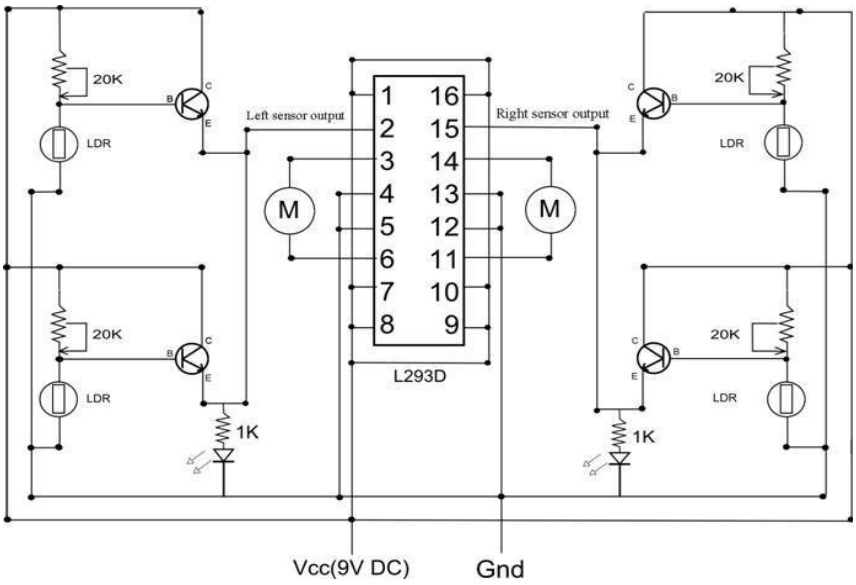

fig. 3 – Circuit diagram for a hard-wired line following robot

The question now arises – are the brains of reptiles and insects also hard-wired – or do they rely on 'software' to make then work?

Now Turing was faced with a machine that employs software at at least four levels. The chess program that he enjoyed playing was probably written in C++ but the program is actually stored in the computer's memory in a partly or wholly compiled form, quite unintelligible to a normal human being. This code is converted into a sequence of numerical instructions at run-time by a resident interpreter which is passed to the CPU where, in all probability, a further hard-wired interpreter breaks down instructions like 'multiply A by B and put the answer in C' into a further sequence of micro-instructions. While Turing was supremely skilled at breaking codes, I do not see how, without a program listing, a C++ manual, a C++ compiler manual and a description of the instruction set of the CPU even he could ever figure out how the machine actually plays chess. Each software level produces an almost impenetrable barrier to upward understanding.

Now while the most recent CPU's contain over a billion transistors, it is often pointed out that the human brain probably contains nearly 100 billion neurons – and if you count *connections* rather than transistors/neurons, the human brain has incomparably more of those. So, the argument goes, if we cannot even understand how computers work (without the necessary manuals, of course) what chance have we to understand even the simplest of brains?

Well, we do understand the simplest of brains quite well, actually. The brain of the nematode worm *Caenorhabditis elegans* has precisely 302 neurons and we know exactly how they are wired up and what they each do. *C. elegans* is basically little more than a line following machine. It is not programmable and has no 'software'. Could it be the case that all unconscious brains are the same? Hard-wired? Non-programmable? If so, how do such creature *learn*? *Can* such creatures learn?

Honey bees can remember the location of a source of food and communicate that information to the rest of the hive when they return so they must be able to learn. Fish too can adapt their behaviour in response to laboratory situations and can learn migration routes from their elders so if we go along with the idea that these creatures are not conscious, we must grant them some means of storing and recalling learnt information. But does this mean that they have 'memory cells' and possess some sort of software 'code' in which to store information? No – it does not. The hard-wired white line following robot can be converted into a black line following robot by swapping round a couple of connections. It does not need a memory – only the ability to adapt its wiring. And this is a feature that we know organic brains possess. There exist remarkable videos of neurons actively seeking out the right connections to make and when human brains grow old and loose the ability to make new connections in their brains, they also loose their ability to make short term memories too.

In this way I envisage even a human brain to be totally hard-wired at every stage of its life. There is no software, no secret code. What you see is all there is. In principle, therefore, I believe that if you constructed a machine made out of silicon with 100 billion transistors all wired exactly like my brain, it would behave exactly like me – but in a coma! If it was connected up to my body it could regulate my breathing, react to light and touch, it might even cause my body to get up and make a cup of tea – but it wouldn't talk sense, it wouldn't compose music or tell jokes. In short, it wouldn't be conscious; it wouldn't be *me*.

So the big mystery of the unconscious brain is not how it works. We already know how it works, in principle at any rate. But two mysteries remain.

The first relates not to how a simple brain *works* but to how it *develops*. How does it know what connections to make as it grows? What general principles govern its morphology and genesis? Some evidence from embryological studies suggest that chemical gradients play a significant role in the early development of a creature's nervous system but can such a crude principle really explain how the complexity of the adult human brain comes about? How do neurons 'know' which connections to make and which to avoid? No-one knows.

The second problem which we urgently need to address is the problem of how memories are stored in the brain and whether short and long term memories are stored in the same way or by different mechanisms. Are there neurons in the brain wired up like the twin transistors in a flip-flop which store information dynamically or are all memories hard-wired? Is learned *behaviour* stored in the same way as learned *information*? How much can an unconscious brain such as that of an insect or a goldfish remember or learn? Apparently the world's cleverest goldfish has learned 9 tricks; how much can you teach an octopus or a fly? Again, no-one knows the answers to these fascinating questions.

The mystery of selfhood

If you can bring yourself to accept the hypothesis that all brains are basically hard-wired robots but that the brains of some higher-order animals can work in a second non-classical mode which gives rise to an emergent property which we call consciousness, then pretty well all the paradoxes which have puzzled philosophers and lay persons down the ages simply evaporate.

We can see how it is possible for the same brain to be in different states of consciousness at different times and how different brains can have different degrees of consciousness. We can be absolutely certain that computers and many primitive brains are not conscious and we have, in principle at any rate, a foolproof test of whether a particular brain is capable of conscious thought or not. This should eventually be able to give us scientifically reliable answers to the questions of which other species of animals are capable of conscious thought and, most importantly, at what stage in the development of the human embryo this capacity emerges. We can see how the facility of conscious thought gives those higher-order animals an evolutionary advantage over their competitors and, from the fact that conscious thought seems to have evolved independently in two separate classes of animals, we can be confident that consciousness is not a miracle or a freak occurrence, but a natural consequence of the laws of physics (some of which are as yet unknown). We can even envisage a time when, those laws being made clear to us, we could (if we were foolish enough) build a conscious being out of silicon or some other inorganic material which would be able to compose symphonies, tell jokes, and even (horrible thought!) exercise its own free will to make war on mankind if it so chose to do so.

So have I answered all the questions which I posed at the start of this essay? Not quite. There is one outstanding issue left which, ultimately goes to the heart of the paradox of consciousness. Put simply it is this: *we only know we are conscious because we are conscious*. On the one hand this sentence appears to be self-evidently true but looked at another way, it appears to be as meaningless as the statement 'oranges are orange because they are orange.' The circularity inherent in our faith in our own consciousness is an indication that consciousness itself (as opposed to a non-classical process going on inside a brain) is not amenable to objective scientific analysis. Consciousness is a meta-phenomenon, outside the realm of scientific enquiry. Even if and when we come to understand the non-classical process that give rise to consciousness, we will still not *understand* consciousness in any meaningfully objective way.

So perhaps we should simply ditch the notion of consciousness altogether. Once we understand the workings of the human brain and can construct non-classical robots to compose symphonies for us, perhaps we should simply admit that that is all there is to it and that our subjective experiences of our own conscious thoughts are at best irrelevant, at worst an illusion. Perhaps we are all, if not mind*less*, then merely mind*ful* robots going about our business in accordance with the non-classical laws of physics, exercising our free will, loving and hating, telling jokes and making war,

thinking irrelevant thoughts and wondering why we bother.

For many, such a soulless scenario is an abomination and I can see why many people will want to reject it out of hand. I, too, reject it but for a different reason. To see why, let us go right back to first principles.

In 1637, René Descartes published his Discourse on Method which contained the famous argument '*Je pense, donc je suis*.' or in English '*I think therefore I am*.' (It was only later translated into the more famous Latin '*Cogito ergo sum*'.) Descartes argued very persuasively that he could doubt absolutely everything else, but the one thing that he could not doubt was his own existence – because he knew he was thinking. In other words, because he was conscious. This argument has become so familiar that it is accepted almost without question and it can be said to be the foundation stone of every Western philosophy ever since. Indeed, if you just sit quietly and contemplate your own thoughts, you will quickly convince yourself that it is true. If '*I* ' am thinking then surely '*I* ' must exist – otherwise what or who is doing the thinking?

But with issues as important as this, we must tread very carefully indeed. I am reminded of the story about the grandfather, the father and the son who went for a trip in a train. The son was looking out of the window and suddenly exclaimed “Look, Daddy. All the sheep in that field are black!”. The father, who was a careful man said: “Well you can't be sure of that. The best you can say is that all the sheep in the field *which you can see* are black.” The grandfather (who was a mathematician) corrected him. “Actually the best you can say is that all the sheep in the field which you can see are black *on at least one side*.”. Let us apply the same rigour to Descartes' famous phrase.

The first criticism we can make is that as soon as we have uttered the first word (Je or I) we have begged the question. This defect is easily remedied as follows. 'Thoughts exist, therefore there must exist an '*I* ' which is having those thoughts.' For the sake of clarity, let us define a 'self' as that entity which has thoughts. We now have: 'Thoughts exist therefore selves exist.' I hope we can all accept the argument so far. '*I* ' have thoughts so I am a 'self'. Presumably you also have thoughts so presumably you too are a 'self'. Since we believe that there are many billions of conscious creatures on this planet all having different thoughts, there must be many billions of different selves.

So far so good. But it is here that we begin to make unwarranted assumptions. When I go to sleep, I lose consciousness. And when I wake up the next morning I regain consciousness and take up my life apparently where I left it the night before. It is natural to assume that the 'self' which wakes up in the morning is the same 'self' that went to sleep the night before. But this does not follow from Descartes' aphorism. To be strictly accurate, the best we can say is “Thoughts exist therefore selves exist *while the thoughts exist*.” We cannot deduce that selves have an independent existence separate from the thoughts which the selves have.

'Oh but this is ridiculous!' I hear you say. 'When I go to sleep, it is still *me* who wakes

up in the morning – not some other self! That would be absurd!'

With the greatest respect, it would not be absurd at all. If a your 'self' were magically to occupy *my* body in the morning, it would remember that it had enjoyed the concert that *my* body went to last night and that it had a dentist appointment that *my* wife had arranged for me at 2 o'clock and that *my* Auntie Mabel was coming to tea etc. etc. In short – it would not be 'you' it would be 'me'. Similarly, if my 'self' woke up inside someone else's body, it would not have any recollection of it being 'me' the day before. It would simply *be* someone else.

In fact, it is the assumption that selves have an independent existence separate from the body it occupies which lands us in so much trouble. Consider the Spock paradox.

When Scottie beams Dr Spock up to the starship Enterprise, all his molecules are scrambled, sent up in a plasma beam and reconstituted inside the spaceship. How can we be sure that the beamed-up Spock is really the same as the original? How does the 'self' get transported? When exactly does the 'self' leave the original body and enter the new one?

Worse still, consider the Kryptonite Man paradox. Kryptonite Man is an exact clone of Superman created by making an exact copy of Superman molecule by molecule. Does Kryptonite Man share the same 'self' as Superman or are there two 'selves'? Suppose we create two cloned Supermen, destroying the original in the process. Which clone would claim to be the 'real' Superman?

And if you think that these examples are too contrived and only prove that teleportation and macrocloning are not possible even in principle, then what about the experiences of patients with schizophrenia or those whose brains have been surgically divided into two? What happens to the 'self' of someone who has Alzheimer's disease or who suffers an accident which completely changes their personality? Is it possible for a self to change as a result of a physical accident? How many different selves can inhabit the same body?

As soon as we abandon the idea of a 'self' having an independent existence, all these questions are answered. Every time the non-classical processes inside a conscious brain are switched on, a 'self' comes into being – and when the conscious brain goes to sleep or dies, the 'self' disappears. There is no contradiction in the idea that a patient with a 'split brain' could think two thoughts at once, or that one half of his brain is unaware of what the other half is experiencing. Nor do we need to be surprised that personalities can change over time, sometimes depressingly so. If the chemistry of the brain changes, personalities may change also. Ultimately, we can see that the concept of a 'self' is a completely empty one and that Descartes' sentence is actually nothing more than a definition and no more or less true or meaningful than the statement 'Mistakes exist therefore boojums exist.' (where a boojum is defined as a thing which makes mistakes!)

But, dear reader, I sense that you are still not satisfied. What is it that provides the

essential sense of *continuity* which we all experience when we wake up in the morning, the sense that I am still the same *person* who went to sleep last night. What is it that gives me the unshakable feeling that inside this mortal body there is a real *me* which inhabits it? Wh*y* do I feel as if I inhabit *this* body as opposed to someone else's? Believe me – I am troubled by these questions in exactly the same way as you are. I am not going to write it all off, as many materialist philosophers have done and say that consciousness in nothing but an illusion. But neither do I believe that there is a *me* which is in some way independent of the body which is typing this essay. The thing which actually defines *me* is nothing more than my body, my brain, and all the neuronal interconnections in it which have been forged over the last 60 years and which constitute my memories and my personality. It is this mortal body which provides the continuity between one day and the next and that is all. If you cannot rid yourself of the notion that there must be a 'self' which is thinking today's thoughts, that is ok because that is the way a 'self' is defined. But whether today's 'self' is the same as yesterdays is a meaningless question. Is the telephone number that you use to phone your mother today the same as the number you used yesterday? I don't mean, is it the same number but is it *the same actual number*? The question is pointless. Mathematically the two numbers are identical but it makes no sense to argue that the numbers are the same in any physical sense.

In truth, it is not a 'self' which thinks thoughts, it is a conscious brain operating in non-classical mode. And the brain that wakes up today is, of course, the same as the brain that went to sleep yesterday.

But I am still hearing howls of protest from my exasperated reader: “OK so my brain provides the continuity from one day to the next but you still haven't explained why I feel that I am so *unique*. Why am *I* here, now, reading this and not someone else, over there playing football? When I look round at all the other people around me I can accept that they are all *like* me but they are not ME! *I* am ME and no one else is!”

Yes, I do see your point. But then I feel exactly the same as you do. So does every other conscious being on the planet. But then that is really my point also. We are all equally unique. The only way to avoid this kind of pointless argument is to reject the idea of a 'self' altogether and restate the case in purely physical terms.

Consciousness, I have argued, is a meta-phenomenon which arises naturally when brains (or possibly other structures) use non-classical processes to make free decisions about what to do on the basis of long-term memories of their past experiences. Since we do not yet know enough about the non-classical processes referred to, we are, as yet, unable to construct conscious machines and even if we reach that state of enlightenment, we will still not *understand* why or how the system we have built is conscious. What we do know, however, is that whenever a system is conscious, it has a powerful sense of being a unique individual. This is not an accident or an unnecessary by-product, it is an essential feature of consciousness and it is the reason why conscious individuals can relate to other conscious individuals in a unique way. This ability to recognise oneself and others as individuals is what

enables penguins to find their mates after months of separation at sea; chimps to show their offspring to make tools to raid an ants nest; elephants to mourn over a lost parent and humans to fall in love.

Every time you wake up in the morning, you do so with this overwhelming sense that your are unique and the same person that went to sleep. You are not wrong. But it is not some disembodied 'self' which is unique and which endures from one day to the next – it is your body and your brain which has been endowed with this miraculous capacity of conscious thought.

Look after it because it is all you have got.

Carr Bank: July 2015