An intelligent layman's attempt to understand

Consciousness

and Free Will

To my wife and lifelong companion

Carr Bank, September 2024

9 781326 870584

Table of Contents

Preface

Amazon currently[1] lists over 800 books published in the last 5 years with purport to be about consciousness and the brain. Some of these are written by authors with a mission, whether it is inspired by their religion or other quasi-spiritual convictions; the majority, however, are erudite works of great scholarship written by respected scientists who are genuinely trying to understand consciousness and the way the phenomenon fits in with our current understanding of the way the brain works.

This book is different.

I am not a neuroscientist or a psychologist and I do not pretend that I possess the answers to the mysteries of consciousness and free will but I may as well warn you in advance that if you are of the opinion that human beings are conscious because they possess an immortal soul (spiritualism) or if you believe that all living creatures are conscious in some degree (panpsychism) then you will find little of interest in the following pages. It is my belief that consciousness is a wholly material phenomenon and that the study of consciousness is a legitimate field for scientific study. While science may never be able to tell us all the answers, we can at least hope to understand which creatures are conscious and which are not, and what it is that makes a conscious brain different from an unconscious one. We may, indeed, be able to use this knowledge in the future to construct conscious machines – though whether this would be an advisable thing to do I will leave future generations to decide. I also believe that free will exists and

1 December 2023

that it is intimately related to consciousness in that you cannot have the former without the latter. Both consciousness and free will appear to me to be incompatible with the laws of physics as we currently understand them. It follows that our current understanding of the laws of physics must change. This should neither surprise nor frighten any individual who is genuinely interested in the search for the truth about the world we live in., On the contrary, the prospect of new discoveries awaiting science is as exciting as it is inevitable.

Of the 800 books in Amazon's list I have rad perhaps a dozen of the most recent and important but I cannot say that I have been persuaded by any of them (though I have listed them in a biliography at the end of the book). The standard form is to spend several hundred pages explaining how the brain works – the purpose of which is more to show how clever the author is than to enlighten the reader – and then to produce some pithy catch-phrase like 'integrated information theory' or 'global neuronal workspace' and with a few pages of high-sounding but impenetrable prose to persuade the reader that the author knows all the answers.

As I say – this book is different.

I am not a specialist and you will learn nothing from me about how the brain works. My approach will be from the other end. To deduce the scope and limits of the concepts of consciousness and free will by considering evidence gained from observing the actual behaviour of humans and animals.

This, then, is the justification for yet another book on the subject. It is an intelligent layman's view and while I will not claim that it is free of errors, omission or unconscious prejudices, I hope that it describes a viewpoint which is both rational and honest.

Turing's laptop

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[2] memory, 72 kbytes 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 Wilmslow. "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

2 Cathode Ray Tube

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 did 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,' 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 test the primality of $2^{521} - 1$, a task which had taken the Ferranti a couple of hours to achieve but which the machine accomplished in a few seconds.

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[3] 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

3 Cathode Ray Oscilloscope

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 be 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 the CPU[4] (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 (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 serious doubts as to whether the current program of reverse engineering of the brain will ever give us a true understanding of the nature of *consciousness*. The gap between the objective nature of the workings of a neuron and the subjective nature of our conscious perceptions seems to me to be far too wide. I will, however, leave you to make up your own mind when we have considered some of the relevant facts which have often been overlooked or even deliberately ignored by other authors.

4 Central Processing Unit

1 - Five important observations

We all know what it is like to be conscious. We do not need to define it. It is simply that overwhelming sense that there exists a thing called 'ME' which is at this very moment situated somewhere in space and time having experiences and thinking thoughts. The subjective experience of seeing a colour or hearing a sound is called *sentience* and the experience of thinking thoughts is called *sapience*. There is some debate[5] about whether a creature can be sentient without being sapient or vice versa but I shall not comment on this except to say that in my experience (and presumably yours also) you cannot have one without the other. So when I talk about a *conscious* creature, I shall assume that the creature is not only having sentient experiences, but is also in some sense able to think thoughts about those experiences and that this gives the creature some kind of sense of *being*, a sense of *self*. Things get a bit problematical when we consider creatures like octopuses whose brains and life styles are so far removed from ours that it is almost impossible for us to conceive what it might be like to be such a creature – but, if a creature has a sense of self, then it must be conscious.

Note, however, that I am in no way implying here that a sentient creature necessarily has to be *self-aware*; only that it has to have a feeling that there is a self inside it which is capable of experiencing sensations and thinking primitive thoughts. I shall have more to say about self-awareness later but for the moment, I shall allow the possibility that a creature can be conscious without

5 See, for example chapter 1 of Nicholas Humphrey's excellent book *Sentience*.

being aware of its own conscious state. Note also that I am assuming that there may well be creatures out there which are capable of sensing and responding to outside stimuli but which have no conscious feelings at all and therefore no sense of self. Such creatures can be said to be *sensitive* but not *sentient*. Even plants are *sensitive* in that they respond to to light and temperature in appropriate ways. Indeed, it is one of the main characteristics of all living creatures that they must be able to sense and respond to changes in their environment. (This is why we classify bacteria as living creatures but viruses as merely self reproducing molecules.)

The big question is – how can we recognise the difference between a creature which is sentient and a creature which is merely sensitive? How can we detect the presence of this internal sense of self in other creatures and what exactly is going on in a creature's brain which makes it sentient and therefore conscious?

Let us start with five important facts about potentially conscious brains, each of which raises important questions of its own.

1) *Potentially conscious 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. Are there certain specialised *structures* within the brain which are only involved in conscious thought? Or is it that the brain is working in a different *mode* when it is conscious? Is the transition between the wakeful state and the unconscious state a gradual one, or is it more of an all or nothing process? Is it right to put all three unconscious states (i.e. being asleep, being under anaesthetic or being in a coma) into the same category, or are they just as different from each other as they are from the state of bing awake?

I shall argue that consciousness arises when certain unique processes occur in the brain and that there is a sharp divide between the conscious and unconscious states which cannot be bridged by appealing to the currently fashionable concepts of chaos, complexity and emergence.

2) *Conscious brains can do things which unconscious brains cannot do.*

This is surely true because, if this were not the case, there would be no evolutionary advantage in being consciousness. Notwithstanding this argument, we still need to establish exactly what it is that conscious brains can do which unconscious brains cannot.

The more we learn about animal behaviour, the more remarkable it appears to be. We are told that dung beetles can navigate by the stars[6] and that stingrays can perform simple addition and subtraction using the numbers one to five[7] but I do not believe that these facts prove that either dung beetles or sting rays are conscious (though I do not rule out that possibility). We know for a fact that it is possible to program an autonomous device such as a cruise missile or a Mars rover to act in a very intelligent way so consciousness must have evolved for a very different reason. I shall argue that the most important advantage which consciousness confers on an animal is the ability to form social bonds with other individuals of the same species.

3) *Conscious beings possess the ability to memorise extremely complex information (including images and sounds) for long periods of time.*

It is widely assumed that memories are stored in the brain by altering the strength of the synapses which transmit impulses from one neuron to another. No doubt this is true up to a point and it is

6 https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0079
7 Scientific Reports, 2022; 12 (1) DOI: 10.1038/s41598-022-07552-2

almost certainly true of our ability to memories motor skills such as my ability to play 'Fur Elise' (after a fashion) 50 years later. But does this theory really explain how I can suddenly remember having been to a certain location before in my childhood or how a doctor can assimilate and remember huge quantities of information about diseases and their remedies?

It seems to me that there is probably a profound connection between long-term memory and consciousness and that we will never understand one without understanding the other. I shall also argue that, although it would appear to be technically possible for a creature to be conscious without long-term memory, in practice the former is not of the slightest use without the latter. It may also be the case that a few unconscious creatures are capable of memorising things for a long time but I believe that this situation is relatively rare. There are a few fish which form long lasting relationships with a monogamous partner but this is not the same as the ability to remember an event or location after many months or years.

4) *Conscious (human) beings report an intense feeling of 'self', of 'being' and of uniqueness.*

This observation is the primary – indeed, the only – evidence that consciousness exists at all. The fact that this evidence is completely subjective has led many neuroscientists to conclude that consciousness is not a legitimate subject for scientific enquiry. This has led them to the belief that consciousness is some kind of 'illusion' and does not really exist. It is just some kind of 'emergent phenomenon' and is best left to psychologists or even philosophers to discuss.

I cannot imagine a greater mistake. I *know* that consciousness exists and I assume that, in their hearts, every committed neuroscientist also *knows* that it exists. As scientists, it is our bounden duty to incorporate it into our current world-view as best we can; and if it doesn't fit, then so much the worse for our current world-view.

Once we accept that the possession of consciousness is a scientifically testable hypothesis, our first task is to look for evidence of it, or its lack, in other creatures. Second, we must search for the physical mechanisms which underlie it (a process which might involve some radically new physics) and finally we might decide to use our new-found knowledge to create conscious machines.

5) *Conscious (human) 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. As we shall see, I think this observation is of crucial importance and must not be ignored, but for now we shall concentrate on the first three observations each of which is telling us something essential about the conscious mind. As we consider each observation in turn in more detail I shall attempt to shed light on the following central questions to which, in my view, any theory of the mind should provide an answer:

Seven important questions

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 the existence or otherwise free-will?

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

2 – Conscious brains

Is here anything about the brain itself which will enable us to distinguish a brain which has the potential to be conscious from one which does not? The most obvious thing which distinguishes the brain of Homo sapiens from those of other species is its huge mass in relation to the mass of our own bodies (though it is not the largest in this respect).

Brain size

The human brain weighs about 1.3 kg or 2% of our body weight. Perhaps even more significantly it consumes about 20% of our daily energy requirements making it an extremely expensive organ to run and maintain. How do these figures compare with some other species?

Most mammals clock in with a brain/body mass ratio of between 0.5 and 1.5 %. Some species of monkey (though, not of our nearest relatives, the great apes) have ratios close to 3% but the brainiest mammal of all would appear to be the Etruscan pigmy shrew which checks in at less than 2 g when fully grown and has a brain weighing 10% of its body weight. It is not so much as a Pigmy Shrew – rather it is a shrewd pigmy!

Not surprisingly, elephants, with a brain/body mass ratio of around 0.02%, do not do well on this measure and neither do whales – which only goes to show that brain/body mass ratio is not a good guide to cognitive ability. However, just for comparison's sake here are some more rough figures for a few other species:

Reptiles	0.01%
Sharks	0.04%
Small fish	0.50%
Dolphins	2.00%
Small birds	8.00%
Ants	14.00%

Perhaps we can get a better idea of the potential for a brain to support consciousness by counting the number of neurons it contains. Fig 1 compares three birds and three mammals with similar sized brains. (The figure at the bottom right of each image is the number of 'cortical' neurons[8] in millions.)

8 Strictly speaking, birds do not have a cortex but similar structures have been identified in what is known as the dorsal ventricular ridge which seem to perform the equivalent function.

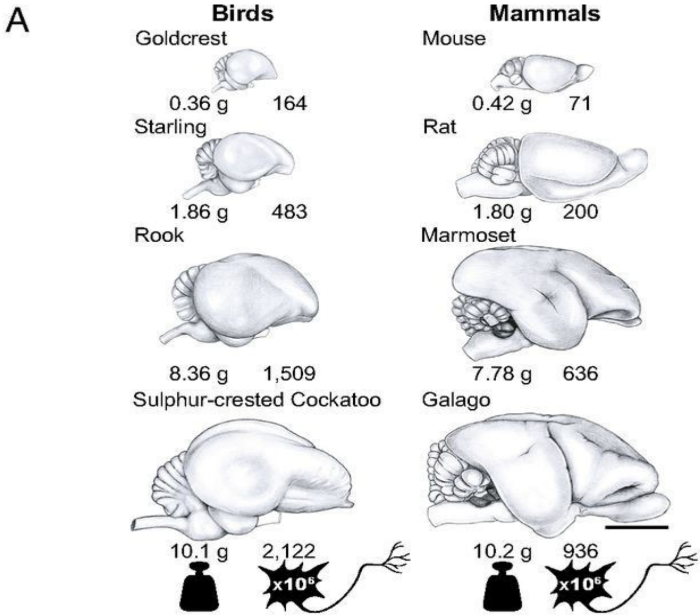


Fig 1: https://doi.org/10.1073/pnas.1517131113

It can be seen at once that birds have between 2 and 3 times as many cortical neurons as the equivalent mammal. This should not surprise us as birds need to keep their brains as light as possible so it makes sense that evolution has found a way to pack even more processing power into the same weight of cerebral matter. So while small mammals pack around 100 million neurons per gram, birds manage up to 250 million neurons per gram. The human brain has 86 billion neurons (of which only 16 billion are cortical neurons) with a density of only 50 million neurons per gram.

Now trying to quantify the 'intelligence' of these six animals is obviously impossible, but I do not think that many people would disagree with the general conclusion that the cockatoo is 'more intelligent' than the rook and that the rook is 'more intelligent' than the starling and the goldcrest. Similarly with the galago (bush baby), marmoset, rat and mouse. Making comparisons across classes is even more controversial but, broadly speaking, there does seem to be a consistent correlation between general intelligence and the number of cortical neurons.

Now let us look at some other classes of animal. Since we know even less about the structure and function of reptilian, crustacean and insect brains as we do about avian and mammalian brains, the best we can do at the moment is count the total number of neurons. For example, a crocodile has a brain whose mass is about 8 grams and contains a total of about 80 million neurons at a density of 10 million neurons per gram. It has fewer neurons than a quail and about the same number as a mouse.

A frog boasts 16 million neurons in total in a brain weighing third of a gram, far fewer than any bird or mammal. A lobster has a brain the size of a grain of rice containing only 100,000 neurons. This is smaller than the brain of a cockroach which contains a million neurons.

Weight for weight, an ant has the largest brain of any animal and with a quarter of a million neurons, it has the potential to be a good deal more intelligent than a lobster.

In a list of animals on the Wikipedia website[9] ranked in order of the total number of neurons, insects clock in at between 0.1 and 1.2 million neurons with the Californian carpenter bee at the top; reptiles have between 1.7 and 80 million neurons; birds between 140 million and 3 billion neurons and mammals between 35 million and 260 billion (the African Elephant).

Sea creatures are not included in the Wikipedia list but zebra fish get along perfectly well with fewer than 80,000 neurons. The great majority of crustaceans, insects and fish have less than 1 million neurons. These data are summarised in Fig. 2[10].

---

9 https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons

10 Note that the horizontal scale is logarithmic. Each unit on the scale

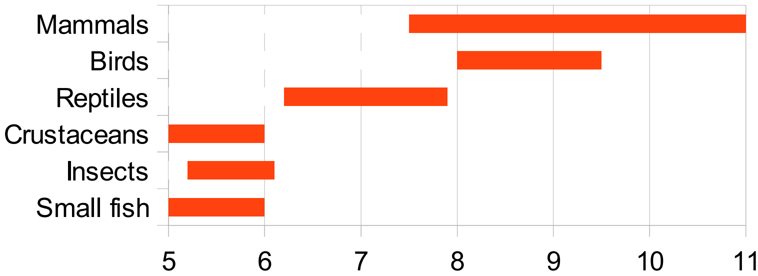


Fig 2: Number of neurons (Log scale)

It is immediately apparent that birds and mammals are in a class of their own. Even the humblest bird has more neurons than the mightiest reptile.

Three important omissions should be mentioned. First, the manta ray has a brain mass of 200 g containing, perhaps, 2000 million neurons (9.3 on the scale and about the same as a cockatoo or small mammal) but this is quite exceptional for a fish. There is some anecdotal evidence that manta rays can exhibit behaviour more typical of a conscious creature and they may even be able to recognise themselves in a mirror but a lot more research is needed here. Sharks, too have large brains but, unlike manta rays, they do not appear to be particularly intelligent.

The other omission is, of course, the octopus which has around 500 million neurons (8.7 on the scale on the chart). Mind you, it doesn't have the sort of centralised nervous system which we would call a brain but the question of whether cephalopods are or are not conscious is definitely wide open.

Is this data sufficient to infer that mammals and birds are conscious but reptiles, crustaceans, insects and fish are not? By no means. But is is at least suggestive. Consciousness has its

---

represents a tenfold increase in the number of neurons. 6 represents 1 million neurons (10^6), 9 represents 1 billion neurons (10^9) etc.

downsides and brains consume a lot of energy so I think we must assume that evolution will only support consciousness if it is definitely to the advantage of the individual and that if we can explain a creature's life style without having to rely on consciousness, then in all probability, the creature will not be conscious.

On the other hand, we do know of one species (which registers 10.9 on the scale of Fig 2) which is conscious so it would seem a little perverse to deny other animals with the same number of neurons at least the potential to be conscious too.

Physiological differences

When we are asleep, we are not conscious. When we are anaesthetized we are not conscious. When we are in a coma we are not conscious. On the other hand, when we are asleep, anaesthetized or in a coma the brain is not dead. Far from it. Even when we are unconscious the brain is monitoring your breathing, heart rate, blood pressure, digestion and much else besides. Is there any way we can tell the difference between a conscious and an unconscious brain by scanning it with some kind of machine?

Most of the activity which is carried out by the brain when we are asleep or otherwise unconscious is carried out in the brain stem – the most ancient part of the brain which joins the spinal cord to the cerebrum. A particularly important organ in the brain stem called the reticular formation is responsible for regulating the body's state of awareness and may even play an essential role in determining whether the brain is conscious or not.

Exactly what the rest of the brain is doing when we are unconscious is unclear, however. When we are asleep, a loud noise may wake us, so the brain stem is allowing some sensory inputs to reach the cerebrum; but patients in a coma or under a general anaesthetic are generally totally unresponsive to sensory stimulation. On the other hand, there is plenty of electrical activity going on even in the brains of comatose patients, but nobody really knows what the unconscious cerebrum is actually doing. Some people think it is consolidating its synapses, some think it is reorganising its filing systems and decluttering its memory banks; maybe there are times when it is doing nothing at all, like a car engine idling in neutral. Whatever it is doing, it is clear that the mere possession of a working brain is not, in itself, sufficient to guarantee consciousness. On the other hand, there must be some difference between a conscious and an unconscious brain – but what is it?

Huge strides have been taken towards answering this question in the last decade, largely due to a plethora of new ways of scanning the brain and vastly improved old ways. The electroencephalograph or EEG has been around for decades but whereas the old instruments used a dozen or so electrodes, modern versions can use up to 256. Then came the magneto-encephalograph or MEG which does a similar job but records the minute changes in magnetic fields that occur when neurons fire.

Newer techniques include functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET) and functional Ultrasound Scanning (fUS). In addition, where circumstances permit, much information has been gained from patients who have had electrodes implanted directly into the brain (for example, to control epilepsy).

The upshot of many experiments involving conscious patients being asked questions and performing tasks while being scanned by one or more of the above machines has been the identification of several recognisable and repeatable patterns of brain activity which are exclusively associated with conscious thoughts. These patterns are known to neuroscientists as the neural correlates of consciousness or NCC's.

In one important early experiment, subjects were presented with words flashed on a screen for a few milliseconds mixed with some random patterns. At this level the words only enter the conscious mind about 50% of the time. Electrodes placed near the frontal lobes of the subject showed a marked spike about 300 milliseconds after the word was displayed *but only when the subject recorded seeing and recognising the word*.[11] Here was the first indication of a pattern of brain activity which is unequivocally connected to a conscious event. But does the conscious thought cause the spike – or does the spike cause the thought?

To answer this question we must stimulate the brain and see if we can generate thoughts artificially. This can be done quite harmlessly in a willing subject using a technique called Transcranial Magnetic Stimulation or TMS. Using electrical coils positioned on the head, a pulse of current can be induced in precise locations in the brain. When the visual cortex is stimulated in this way, flashes of light are seen.

This still does not resolve the issue, however, because stimulating the visual cortex is not the same as stimulating the neuron which supposedly is the one that causes the conscious thought. It has even been shown that there exist neurons in the brain which only fire when the subject is consciously thinking of a very specific subject such as the Sydney Opera House. It does not matter if the subject is shown a picture or the words, the same neuron fires; and if the neuron is stimulated artificially, the subject will report thinking of the object. But once again, we have only demonstrated a close correlation, not a necessary causal effect.

In spite of the experiments described above, many neuroscientists are sceptical of the claim that there is, as it were, a one-to-one relationship between individual neurons and even primitive conscious thoughts. More attractive to many is the idea

11 This experiment is described in detail on page 124 of *Consciousness and the Brain* by Stanislas Dehaene

that conscious thoughts relate to a *pattern* of activity, not just to the firing of a few individual neurons. According to these scientists, if you want to determine whether a brain is or is not conscious at a particular time, you must examine the whole brain, not just isolated parts of it.

The first experiments along these lines were carried out by Massimini and Tononi in the first decade of this century. They used TMS to stimulate a region of the brain and then, instead of asking the subject what they thought, they used a sophisticated EEG machine to record the whole brain's response to the stimulus. Then, having gathered a large amount of data from each of the 256 electrodes, instead of analysing the data in detail, all they did was to compress the data using a standard compression algorithm just like the one that compresses the 10 megapixel images you take with your phone into a 1 or 2 megabyte jpeg file. What they found was truly amazing. In conscious patients, the data was highly complex and it was not possible to compress it very much. But with unconscious patients the algorithm found many repetitive elements and was able to reduce the size of the data considerably more. Using a suitable mathematical algorithm it proved possible to generate a single number called the PCI[12] which gave a pretty reliable indication of whether the subject was or was not conscious.

Since then, techniques have become more and more sophisticated and we are currently on the verge of using some of these techniques in the hospital, rather than the lab, to inform life and death decisions as to what to do with patients who have suffered massive trauma and who appear to be unconscious. The level of consciousness of such patients is often assessed by grading their responsiveness to various degrees of stimulus according to an internationally agreed scale such as the Grady Coma Scale. But behavioural tests like this can be misleading.

12 Perturbation Complexity Index

Clinicians have learned to distinguish wakefulness or arousal from awareness or consciousness. Normally, levels of wakefulness and levels of consciousness change in parallel, which is why wakefulness is usually a good enough guide to a subject's level of consciousness. But sometimes a patient suffering severe trauma can be apparently awake but quite unconscious. This state is known as 'unresponsive wakefulness syndrome'[13]. An even more terrifying state is known as 'locked in syndrome'. Here the patient is paralysed to such a degree that he or she may only be able to move their eyes. Nevertheless, once communication has been established it is clear that these patients are perfectly conscious even though they are totally unable to respond to stimuli. One such patient, Jean-Dominique Bauby was even able to write a book – *The Diving Bell and the Butterfly* – describing his awful predicament.

Patients in a coma or under a general anaesthetic are believed to be totally unconscious and therefore oblivious to pain. But how do we know this? Could it be possible that when we are undergoing open heart surgery we feel every incision of the knife and every snip of the scissors? Of course, when we are brought round, we have no memory of these distressing events but maybe the effect of the anaesthetic is not to shut down our conscious perception but only to wipe clean our memories of them.

Fortunately, I think we can discount this possibility. What non-invasive techniques we currently possess for detecting consciousness indicate that the brains of patients in a coma or under a general anaesthetic do not respond in any way that would indicate that they are responding to what would normally be regarded as painful stimuli. In addition, it seems very unlikely that anaesthetic drugs could disable our memory systems but leave our subjective perceptions intact. Conscious perceptions and our memories of conscious perceptions are probably two sides of the

13 also known unhelpfully as the 'vegetative state'

same coin and if you disable one, you automatically disable the other as well.

So where does all this leave us?

Obviously there is a difference between a conscious brain and an unconscious brain and we can sometimes tell the difference by monitoring the brain's electrical activity. It is widely assumed by the scientific community that all conscious thoughts are caused by, or at least directly correlated with, certain patterns of activity in the brain and we are beginning to identify some of these neural correlates of consciousness but we are still a long way from understanding why the patterns of activity bring about the phenomenon of consciousness. Current diagnostic tests for consciousness fall far short of being able to tell us whether consciousness exists in other animals and it seems unlikely that studies of the neural correlates of consciousness will enable us to program conscious computers any time soon.

The upshot is that, while the neurological study of the brain is a tremendously exciting field and is likely to greatly improve our understanding of how the brain works and how better to treat disorders of the brain, I do not, personally believe that neuroscience is going to answer any of the questions posed on page 13 except possibly the first.

In order to make further advances we must turn to behavioural studies of consciousness, the most obvious of which is the study of sleep.

Sleep

Almost as soon as the electroencephalograph (EEG) was invented it became clear that up to 5 different characteristic wave patterns can be observed in the human brain while awake and asleep (see fig. 3).

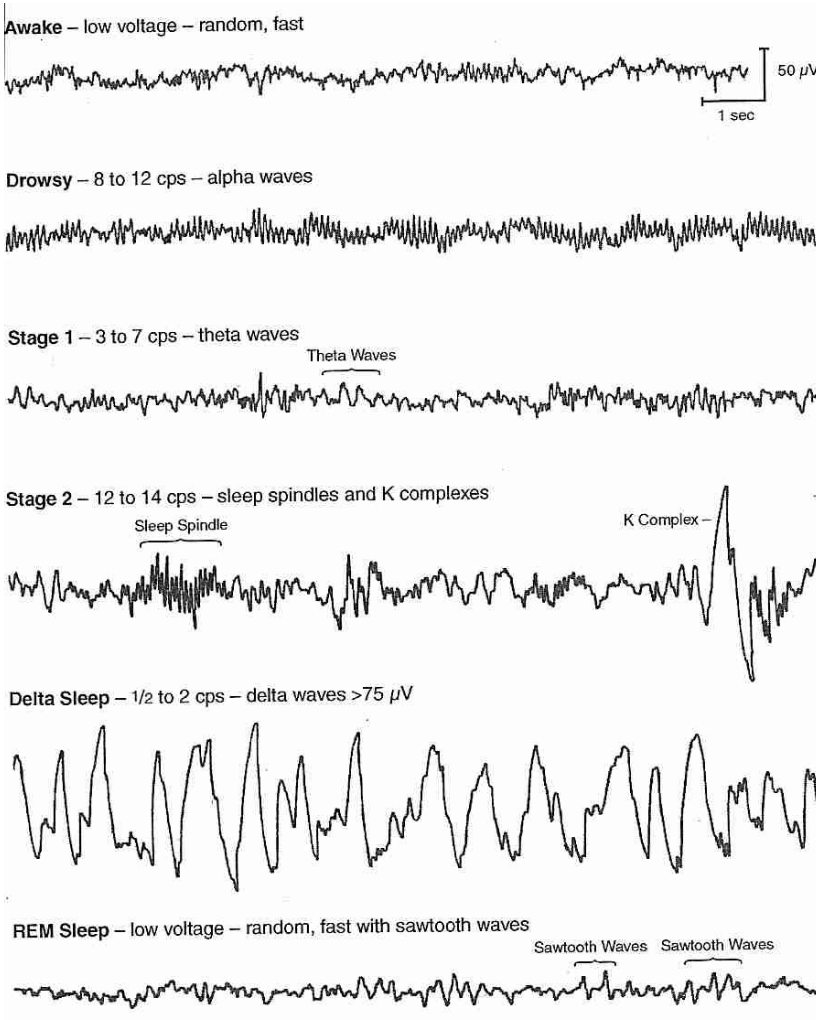


Fig 3: EEG wave patterns (from http://www.benbest.com/health/sleep.html)

When the subject is awake and conscious, the EEG records rapid and irregular oscillations of relatively smallamplitude. 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' and that the information content of all this activity is very high.

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 significantly, 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 eyes 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*?

If your reaction is to say “Of course I wasn't conscious – I was asleep!” I ask you to remember an 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? And when you recall the dream, do you also recall the fact that you were unconscious at the time? No you do not. The recollection of the dream is almost as vivid as the recollection of a real experience.

All the physiological signs indicate that we are indeed as conscious during REM sleep as we are when we are awake; the EEG machine records signals typical of full wakefulness and the fMRI machine shows that oxygen levels in the brain are significantly increased. The only difference between REM sleep and being awake is 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). The only puzzle is why we generally remember so little of what is going on in our brains during REM sleep.

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.

There are many stories about people who have solved problems or had great ideas while asleep. Very often we ascribe these insights to our unconscious mind but I suspect that these ideas may actually come to us when we are in REM sleep. We may think that we solved the problem unconsciously but that is only because we do not remember solving the problem consciously. If, as most people believe, it is only the conscious mind which is capable of thinking in abstract, linguistic or mathematical terms, then I think it likely that the poet who wakes one morning and hastily scribbles down a new poem and the mathematician who suddenly sees the way to the proof of a new theorem has actually worked it all out during REM sleep.

What about sleepwalkers; are they conscious? Are they acting out their dreams? The answer to this question is unambiguous. Their EEG waves are those of a person in deep sleep. Although sleepwalkers have been known to carry out highly complicated procedures, even to the extent of getting into the family car and driving to work, they are generally unresponsive and fail to recognise people they know. Although their eyes are open and they can obviously take action to avoid major obstacles, their judgement is impaired and they often cause injury to themselves, especially when encountering the unexpected. In fact they show exactly the sort of behaviour which we might expect from an unconscious but pre-programmed creature like a prototype robot or an insect[14].

If, therefore, we accept that we *are* conscious during 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 (e.g. 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

14 I ought not to reveal one of my prejudices here so if you are of the opinion that insects are conscious, please ignore this simile and read on.

sleep with only half of their brains at one time[15]! 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. It is said that many 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.

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. Apparently fish also spend periods of time in which they appear to be asleep, but their brain activity is difficult to record so whether their brain activity changes in the same way that it does in mammals, birds and reptiles is largely unknown.

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 show 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[16]. If this is true, it raises an interesting question with regard to the evolutionary development

---

15 This is called 'unihemispheric sleep'.

16 Apparently dolphins do not show any evidence of REM sleep. This could be taken to mean that while the existence of REM sleep may be a reliable indicator for the *presence* of consciousness, its absence does not imply that, for example, reptiles and insects are incapable of conscious thought.

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 either consciousness has evolved separately in mammals and birds or that reptiles have lost that facility. 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.)

Pain

There is another way in which we can judge whether a person or an animal 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.

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. But their response is greatly muted or even completely suppressed when fast asleep or anaesthetised.

Surprisingly little is known about how general anaesthetics actually work. Local anaesthetics such as the ones dentists use simply temporarily disable the sensory nerves leading from the tooth to the brain. Regional anaesthetics such as the epidural given to women in labour can disable the nerves for a whole region of the body. In principle it would be possible to inject a regional anaesthetic into the brain stem thereby rendering the patient insensitive to all sensory inputs; but this would not be the same as being under a general anaesthetic. If a general anaesthetic simply blocked all nervous action, the patient would die. Somehow, the chemical must block those neurons which are responsible for conscious thought while leaving other essential sensory and motor neurons intact. EEG traces obtained from patients under varying degrees of general anaesthesia show similar features as those from sleeping patients so, if only we knew how these agents caused us to go to sleep, we might learn a lot about what makes us conscious.

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, therefore, that all mammals can experience pain and we can confidently conclude that 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 mammals, birds are not capable of suffering pain.

When it comes to fish, reptiles and other animals, we should be very careful to define exactly what we mean by 'pain'. In my book, pain is a *conscious*, *subjective* experience just like seeing a red poppy or experiencing a loud noise. The reason why we do not experience pain when under a general anaesthetic is not because the sensory nerves which are cut by the surgeons knife are blocked from sending signals to the brain but because we are unconscious. If you claim that fish feel pain when they are hooked on a line or that crabs feel pain when they are thrown into boiling water, then you are tacitly assuming that they are conscious. If fish and crabs are unconscious creatures, then, by definition, they cannot feel pain.

The subjective feeling of pain should not be confused with the normal physiological and neurological reactions to painful stimuli. When you lift up a hot casserole dish, special sensors in your hand called nociceptors send signals towards the brain telling you how hot the dish is. If the signal is strong enough, they may trigger an automatic and involuntary reflex in the brain stem which causes you to immediately withdraw your hand. If the signals are not that strong they pass up to the higher levels of the brain which may then make a conscious decision as to whether or not to bear the pain and carry the dish to the table. In either case, other physiological reactions may occur such as increased heart rate, rapid breathing and the production of hormones such as endorphin which help deaden the pain.

Now when fish are treated to a, presumably, painful stimulus such as injecting their lips with acetic acid, they too show similar signs of being in pain. There is a burst of electrical activity in all parts of the brain, increased rate of breathing and the release of opioids such as endorphin. The fish also behave differently, rubbing their lips along the side of the tank and forgoing other normal behaviour such as feeding until the stimulus wears off. All of this is consistent with the idea that fish consciously feel pain – but I am afraid it falls far short of being proof.

It is obvious that the signals sent from the nociceptors to the brain have a vital role to play in the survival and well-being of the individual creature. Without these signals the creature would never learn to avoid the potentially life-threatening situations which the nociceptors are designed to detect. The difference, however, between the fish and the human is that the (conscious) human has the option of overriding the signals from the nociceptors but the (assumed unconscious) fish cannot.

It is an interesting fact that the great majority of insects do not have nociceptors and I think we can confidently conclude that if you pull the leg off a Crane fly (also known as a Daddy long-legs) you are not causing the insect any pain. On the other hand, nociceptors have been found in some worms and molluscs and almost all creatures including insects are capable of generating opioids.

The implication of these facts is that, if you are going to regard the presence of nociceptors, the production of opioids and the existence of behavioural changes as clear evidence of subjective pain, then you are going to have to ascribe consciousness in some degree to virtually all of the animal kingdom. This is, of course, a perfectly logical position to adopt, but I cannot help feeling that this conclusion flies in the face of the general principle that Nature only supplies characteristics to animals on the basis of necessity. If a worm, or even a fish, has no need of consciousness, than I think it unlikely that evolution will have provided it.

It seems to me to be far more reasonable to regard all these reactions to potentially harmful stimuli as mere pre-programmed unconscious responses which have come about through evolution simply to maximise an individual's chances of surviving to reproduce.

The answer to the question 'Do fish feel pain?' is exactly the same as the answer to the question 'Are fish conscious?'. You cannot answer the latter questions just by looking at the behaviour of a few specialised neurons or some stereotypical behaviour. The question is far more subtle than that and we must first address the question of what consciousness is *for* – i.e. what evolutionary advantage it confers on those creatures that possess it.

3 – *What is consciousness for?*

All creatures living today are the result of millions of years of Darwinian evolution. The diversity of all these creatures is truly staggering. Watching a television programme on life in the deep ocean, you might be forgiven for thinking that these creatures lived on a different planet.

But all these creatures have one thing in common. It is a cliché to say it but they are all uniquely adapted to the environment in which they live. With only a few exceptions which prove the rule, every limb they possess, every behavioural characteristic they display, every sense organ they utilise has a purpose whose ultimate goal is to better enable the individual creature to survive and reproduce.

If some animals are conscious, therefore, I believe that it can only be because being conscious *gives them an advantage over the competition*.

This, then, is the main reason why I am so reluctant to believe that worms, molluscs, amphibians, crustaceans, insects and even fish are conscious: I just don't see why being conscious would give them a competitive advantage in the struggle for life.

In searching for the potential evolutionary benefits of consciousness we must start with the most fundamental aspect of consciousness which is the fact that *all conscious creatures have a sense of self*. As soon as a creature becomes conscious of something – a bright light, a loud sound or whatever – almost by definition this implies a 'self' which is experiencing the light or the sound. Most of the arguments about consciousness centre on the exact nature of this sense of 'self' and I shall have much to say about this in later chapters but for now, whatever our views on the nature of consciousness and the mind, we are forced to adopt the convention that if subjective experiences exist, there must be a subject which experiences them. In particular, I do not wish to imply by using the word that the creature is necessarily *self-aware*.

So how does a sense of self give a creature an evolutionary advantage? Consider one of those creatures which lack consciousness. (I am assuming, dear reader, that if you are a panpsychist, you will have already consigned this book to the waste bin.) For the sake of argument, suppose that crabs are not conscious and have no sense of self. All their actions are, as it were, pre-programmed. When you turn over a rock and find a crab underneath, it instinctively tries to hide. This reaction is hard-wired into the crab's nervous system and does not require that the crab be conscious of the sudden increase in brightness around him.

We must now ask ourselves, what evolutionary advantage would be gained by the crab if it had a sense of self and was actually *conscious* of the light? The answer is, absolutely none. In fact, since conscious reactions in humans are measurably slower than unconscious ones, it would be a positive *disadvantage* to the crab to have to process the information consciously.

I strongly suspect, too, that in order to carry out the same amount of information processing conscious brains have to be much larger than hard-wired brains. If this is true, then the advantages of having a conscious brain will have to be considerable in order to make its acquisition worth while.

In this chapter and elsewhere in this book I shall discuss a number of different aspects of consciousness, many of which may be of evolutionary benefit, but there is one aspect of a creature's behaviour which is often confused with consciousness and that is general intelligence.

Let us be clear from the start. *An animal does not have to be conscious to be intelligent*. In recent years experiments have revealed to us the amazing abilities of a wide range of animals to carry out what appear to be highly sophisticated cognitive tasks and I shall examine some of these in Chapter 4. This has resulted in an explosion of popular books with titles like '*The Mind of a Bee*' and '*The Genius of Birds*' and a fascinating television series called '*Animal Einsteins*' which described, amongst other things, the amazing navigational abilities of octopuses, mantis shrimps an even dung beetles. At the same time, however, there has been an equally remarkable explosion in the capabilities of artificial intelligence software and autonomous vehicles. There is no suggestion that ChatGPT or Google's driverless cars are conscious so the existence of intelligent behaviour or navigational ability in an animal is no way to tell if it is conscious or not.

What, then, is consciousness *for* if it is not to make an animal more intelligent?

I believe that by far the most important characteristic of a conscious individual with a sense of self is the ability to *recognise another member of the same species as another individual*. Philosophers refer to this ability as *theory of mind.* Another phrase which is often used is *relational intelligence* but since this term has been used to describe the behaviour of ants and bees which, I believe, are not conscious, I prefer to use the word *empathy* to describe a characteristic which is exclusive to conscious creatures. Note, however, that empathy, as defined here, should not be confused with 'sympathy'.

Empathy

Apart from the obvious necessity of being able to recognise another creature as a member of the same species – for mating purposes if nothing else – I believe that there are two ways in which being able to recognise other creatures as *individuals*, similar to yourself, can be advantageous from an evolutionary point of view. My contention is that only such animals can a) form societies and b) make long-lasting pair-bonds.

One definition of a social species is as follows: “a group of animals belonging to the same species, and consisting of individuals beyond those in a family unit, who perform specific tasks, spend distinctly more time together, and interact much more within the group than with members of the same species outside of that group.”[17] Obviously pretty well all mammals are covered by this definition, exceptions being solitary animals like polar bears and leopards. The trouble is, from my point of view, that the definition also includes ants and bees which, as I have indicated, I am reluctant to include in the list of conscious animals. In my view there must be something fundamentally different between a colony of ants and, for example, a troop of baboons.

Baboons and other primates live in a highly stratified society with each individual enjoying a well-defined place in the social hierarchy. Each individual is known to the rest of the troop personally and every individual knows how to behave in the presence of, say, the alpha male, or a young virgin female.

Much the same is true of a pack of wolves, a colony of mole rats or a pod of whales. Every individual knows his or her role and the roles of the other members of the society.

Now we can see why ant and bee societies are fundamentally different. Ants and bees have sophisticated mechanisms of communication which enable the colony to carry out complex cooperative tasks like foraging and nest building; I imagine that they can also recognise other individuals as being from the same colony by means of their scent etc. But I see no evidence that ants

17 Encyclopedia.com

or bees can recognise each other *as individuals*. What would be the point anyway? One ant is as good as another really. Ant colonies and bee swarms are not societies in the same sense that a pod of whales or even a herd of deer is a society.

Joining together in a society has a great many benefits for a species. For a predatory species like wolves and orcas, it is the ability to hunt cooperatively in a pack which enables them to bring down prey which a single individual animal could not tackle. For a vulnerable species like the prairie dog or an antelope, it is the ability to have many eyes on the look out for potential danger that gives the society survival value.

The conclusion is clear. If you see a group of animals behaving in a way which suggests that the individuals in it are behaving differently towards other individuals in the group, then the animals clearly have a sense of self and must be conscious. Pretty well all mammals fit this description, the exceptions being those mammals which lead essentially solitary lives.

What about birds such as rooks and sparrows who roost together in large flocks? What about alligators which colonise a swamp? What about frogs in a pond or shoals of fish? Many animals congregate in large groups, usually for protection. But the crucial point is whether they *behave differently towards other individual members of the group*. In other words, *is there evidence of a social hierarchy*?

Recent research has unearthed a surprisingly complex social hierarchy in flocks of vulturine guinea-fowl – a species not otherwise known for its intelligence. But the exceptional nature of this discovery only highlights the fact that the great majority of non-mammalian species who gather in colonies for safety or for breeding purposes do not form societies in the sense I am referring to.

While the alpha male of a troop of baboons uses his powers of empathy to dominate his rivals thereby enabling him to mate with a large number of females, birds have used this capacity of empathy in a completely different way. Instead, they have used it to bond with another individual either for a season or, in some cases, for life, thus ensuring a stable partnership which greatly increases the chances of their young surviving to maturity.

A female osprey, sitting on her familiar nest waiting for her mate to return from West Africa shoos off several potential suitors. Why? She could not do this unless she knew that they were not the reliable individual with whom she had reared several successful broods in previous years.

A male emperor penguin will wait patiently warming the egg containing his future offspring for a whole Antarctic winter in the confident belief that his partner will eventually return with food for them both. And, while we must be careful not to read our own emotions into the scene, it is almost impossible not to interpret the the behaviour of the two birds on their reunion as images of joy and even ecstasy.

It is, I suppose, conceivable that an unconscious creature could be programmed to recognize individuals in a society or partners in the task of rearing young, but we know that, in at least one species, Nature has chosen to do it in a different way and when we see so many parallels to our own behaviour in the behaviour of a jealous chimpanzee or a pair of courting grebes, it is simply perverse to deny these creature a sense of self and an ability to empathise with other members of the same species.[18]

---

18 It may not be necessary to include the phrase 'other members of the same species'. When a herd of antelope is grazing peacefully within 100 m of a pair of resting female lions, is it because they are simply unaware of the danger – or is it because they can actually read the body language of the lions and know that the lions are not hungry and in no mood for a chase?

Imagination

Although I believe that the ability to recognise other animals as individuals is the most important reason why conscious creatures have an evolutionary advantage, consciousness comes with other benefits which may have proven to be equally if not more important in the long run. One of these is *imagination* or *creativity.*

There is a pod of orcas in the Southern Ocean that has developed an entirely novel way of catching seals. Having identified a seal quietly resting on a smallish ice floe, four whales swim in close line abreast formation at speed towards the floe. At the last minute, the four whales dive beneath the floe causing a small tidal wave to wash over the floe, washing the unsuspecting seal into the water where it can easily be caught. This behaviour is relatively new and is used by no more than 100 whales so it cannot be described as instinctive behaviour. One of the whales must have thought of the idea first and tried it on his own, probably unsuccessfully. Another whale must have seen him try it and realized that the idea had potential. Eventually, after many trials, the technique involving the coordinated actions of four whales has been perfected and the idea has passed on to a number of other pods in the area as well. I do not believe any of this could happen without conscious thought.

Cooperative behaviour is not unique to mammals. Sharks and other fish are known to cooperate in foraging for food. Ants and bees obviously cooperate in building nests and hives and there are many pairs of species which gain mutual benefit by behaving cooperatively such as ostriches and zebras teaming up for mutual protection, cattle egrets picking ticks off the hides of buffalo etc. etc. but the behaviour of the orcas described above is totally different. It is not the product of thousands of years of evolution and it is not species wide. It only appeared recently and is restricted to a small fraction of the species as a whole.

It must be admitted, however, that creativity and imagination is rare in the animal kingdom. Primates and some birds are known to use simple tools in the wild and laboratory experiments on rats and pigeons show that they are capable of feats of memory, learning and occasionally imagination. Of course, the one species which developed the art of imagining things and creating new ways of doing things and, for better or worse, enabled it to dominate the whole world is our own – *Homo sapiens*.

Intention

The third exclusively conscious faculty is the ability to *remember what has happened in the past* and to use this information to *plan actions in the future* – an ability which I call *intention*. An obvious example of this is the dog that drops his ball at the feet of his master. It is clear that he remembers the many occasions in the past when his master threw the ball for him to retrieve and eagerly expects his master to do it again. A bower bird will spend hours decorating his bower in the expectation that the coloured beads will attract a suitable mate. The matriarch of a family of elephants will patiently guide her flock to a long-remembered distant water hole and a homing pigeon will use remembered landmarks to find his way back to his roost.

It is possible to argue that all these examples of supposedly intentional behaviour are simply hard-wired into the creature's brain; after, all, birds and mammals are not the only creatures that, for example, make long distance migratory journeys. What about salmon fish and monarch butterflies? Are the adult salmon conscious of the twists and turns of the stream which they swam down as youngsters many years ago as they return to their spawning grounds? I doubt it. It is fairly well established that they are following chemical cues in the water. Do monarch butterflies consciously remember the places where they were born when they return to their summer residences in North America having over wintered in Mexico? Certainly not for the simple reason that the butterflies which return are not the same as the butterflies which left, the former generally being the grandchildren or great grandchildren of the latter!

Nor is it the case that it is only mammals and birds like the bower bird which carry out actions that appear to have long-term goals. Spiders build webs in anticipation of catching flies and male cichlid fish build nest structures in order to attract females. But the difference here is that spiders and fish do not have to *learn* these skills either from their parents or by repeated trial and error – they appear to be able to do these things by instinct. It is simply that evolution has selected for those creatures which build better webs or more attractive nests and these actions are hard-wired into their brains.

It must be admitted, however, that it can be very difficult to say in any specific instance whether an apparently goal-oriented behaviour is or is not evidence of intent. No dog owner, however, will deny that dogs know what they want and have sophisticated strategies to get what they want based on memories of what strategies have worked before.

Predatory animals who hunt cooperatively in packs will also benefit greatly from the ability to remember the past and to use the experience gained to plan the future. We have all watched in fascination at the films of killer whales corralling a bait ball and jackals isolating a young buffalo before moving in for the kill. These techniques have to be passed from generation to generation. Young cubs are often laughably inept at hunting. These skills are not instinctive, they have to be learned and mastered.

What about sharks? They hunt in packs. But they are not mammals, they are fish. Does this mean that sharks are conscious? No open-minded scientist should rule out the possibility but until some intrepid (or possibly foolish) researcher is prepared to carry out the necessary investigations, I will put my money on the shark's behaviour as being instinctive, not learned.

To summarise, I have identified three important ways in which the possession of a conscious brain can be of use to an animal which I call *empathy*, *imagination* and *intention*.

Empathy is the ability to recognise other animals as individuals. It is this that enables social animals to live together cooperatively to their mutual benefit and for animals who pair for life to establish long-lasting relationships.

Imagination is the ability to 'think outside the box' and to create new artefacts or new ways of doing things.

Intention is the ability to use the memory of past experiences to decide what to do next.

If we are going to use these concepts as pointers to decide whether or not a particular species of animal is or is not capable of conscious thought, we must first consider whether it is really true that no unconscious creature or artificial system could display any of these characteristics. In short, we must ask ourselves what are the limits on the behaviour of unconscious creatures?

4 – *Unconscious creatures*

Empathy, imagination and intention. These are the three most important characteristics which, I believe, distinguish conscious from unconscious creatures. Of the three, it is the last which is the most problematic because all creatures, almost by definition, necessarily exhibit what I shall call goal-oriented behaviour. Many creatures achieve their goals by what can often appear to be smart, imaginative, sometimes even devious behaviour and it is all too easy to describe them as being 'intelligent'. Indeed, there is nothing wrong with that as long as we are aware that creatures can be remarkably intelligent without necessarily being conscious. And if we are going to use the word intelligent to describe the behaviour of a non-conscious living organism, then we should not be embarrassed about using the word to describe the behaviour of man-made artefacts such as thermostats and self-driving cars.

Let us take a look at some allegedly 'intelligent' behaviour.

Locomotion

Volvox is a type of algae which consists of a spherical colony of about 50,000 cells. The surface of the sphere is coated with specialised cells which have appendages called flagella and by waving these flagella in a coordinated fashion, the colony can swim about; in particular, it can swim towards the light. Undoubtedly this is 'goal-oriented' behaviour – but the object in question is not even a creature; it is just a collection of independent cells which have decided to organise themselves in a particular way. How the cells coordinate their actions to achieve the desired goal is still a mystery.

Most animals rely on their locomotive abilities to forage for food. But foraging is never a random process. Take the much studied nematode worm *Caenorhabditis elegans* for example. This unassuming creature has a nervous system comprising of exactly 302 neurons which controls its behaviour. When deprived of bacterial food, a worm will move around at random for about 20 minutes and then move off in a straight line. Clearly this is an intelligent strategy and maximises the worm's chance of finding food somewhere. But *Caenorhabditis* does not have anything which we could call a brain. It is simply pre-programmed to perform a task in the same way that a factory robot is programmed to follow a white line painted on the floor.

Bees have a sophisticated foraging strategy which ensures that they visit every flower on a bush once and once only (up to a point, at any rate). Ants forage at random until a source of food is found. But then, the lucky ant returns to the nest leaving a trail of pheromones for other ants to follow. But neither of these strategies requires much more than a piece of clever programming.

The need to be able to get to where you want to go is even greater for predatory species so the ability to coordinate different parts of the body to effect purposeful motion is obviously of prime importance to any mobile animal and it must be expected that a large proportion of the brain of an insect, frog, a crab or a fish is dedicated to the task of walking, flying or swimming. Just what a difficult task this is is born out by the fact that it is only within the last few decades that we have been able to construct autonomous devices capable of doing any of these things.

Navigation

Getting to where you want to go is sometimes only half the story. If you have a burrow or a nest then you must be able to get back there.

Many species of limpet make their home in a particular location on a rock (called a 'scar'). After foraging for food at low tide they consistently return to their scar and re-orient themselves so that they fit snugly on the rock to survive the dashing of the waves as the tide rises. It is a bit of a mystery how they know how to return to the scar. They do not navigate by the stars, nor do they carry a map of the rock around with them. The best guess is that they leave an almost indelible scent trail behind them whenever they leave the scar but this does not explain how they know in which direction they must follow the trail to get back home, nor what happens when the rock they live on is completely covered in trails.

Even more remarkable in its way is the single-cell organism *Physarum polycephalum* (a species of what is rather uncharitably called a 'slime mould') which, it has been claimed, can solve mazes and find optimum solutions to the famously difficult 'travelling salesman' problem. The organism spreads out in search of food and in doing so constructs a network of filaments which connect the different parts of its 'body' together in the most efficient way. For example, when scraps of food were placed on a map of the USA at all the major cities, a slime mould eventually organised itself in a manner which closely resembled the actual network of interstate highways.

Would you describe either the limpet or the slime mould as 'intelligent'. Probably not. Both simply seem to be doing what their biological structure is telling them what to do. It is, however, worth pointing out that it is not at all obvious *how* either of the creatures do what they do and both are objects of active current research.

It has long been known that insects such as bees and beetles have good vision and that they not only use the Sun and Moon as beacons to help them find their way around, they also use polarised light from the sky. Unlike bees, however, dung beetles are active at night. Curious to know how they were able to find their way back to their burrows even on moonless nights, researchers at Lund University in Sweden made the astonishing discovery that dung beetles use the Milky Way as a navigational aid to keep them from wandering from the straight line.

More recently, similar behaviour has been identified in mantis shrimps which forage for food and return to their burrows to consume it. Returning home can, however, be a bit of a challenge. Patel and Cronin[19] have proved that the shrimps maintain a pretty good memory of where they are in relation to their burrow by 'dead reckoning' and that they consistently use the position of the sun and the orientation of polarised skylight to assist them in finding their way back home.

Information Processing

Everybody knows that the Venus flytrap (*Dionaea muscipula*) catches insects by closing its leafy 'jaws' when sensitive hairs on the inside of the leaves are touched by its hapless prey. But did you know that, in order to prevent premature and wasteful reactions to other sorts of vibration, the closing of the jaws is only triggered if the hairs are touched *twice within 20 seconds* of each other? The significance of this is that the response is not just a 'knee-jerk' reflex on the part of the plant. The plant is effectively absorbing and integrating information from a range of sensors and then making a decision as to whether or not to respond. In other words, the plant is *processing information*.

The flytrap's information processing abilities can be compared to that of a sophisticated thermostat which responds to a fall in temperature by operating a switch only when the temperature has fallen below it target level for a minimum length of time.

Neither the plant nor the thermostat possess anything like a nervous system and only the most committed panpsychist would

19 Current Biology 30, R639 (2020)

maintain that plants and thermostats are conscious to any degree, but the currently fashionable theory is that even the human brain is fundamentally just a very complex information processing system so the important question for the modern mind theorist is just how complex does an information system have to be before is can support or, perhaps, becomes conscious?

Lets work up the scale of complexity gradually.

Spiders are pre-programmed to build a certain design of web. They do not learn how to do it from their parents and orb spiders never build tunnel webs. Their web-building expertise may be likened to that of a carpet loom which can manufacture a fabulously complex design given the right information in the form of a series of punched cards or, these days, on a CD.

But granted that web-building ability is hard-wired into the spider's brain, the big question is – how is this information (which is presumably coded in some way inside the spider's DNA) translated into the correct network of neurons which will carry out this intricate task? No one knows. Indeed, while the phenomenon of consciousness may seem utterly mysterious, even miraculous, the whole process of morphogenesis by which single celled eggs turn into mature adult creatures is possibly an even greater mystery.

While web-building may not appear to require much intelligence – only a simple rule of thumb, hunting spiders, on the other hand, do display some remarkably intelligent behaviour. There are well-documented cases where a hunting spider (e.g. *Portia fimbriata*) will mimic the behaviour of, for example, a male fly in order to entice a female out of her nest. In another experiment, a spider appeared to select the correct route to some food even though that route initially involved moving *away* from the food. It has even been suggested that spiders can show 'surprise' when presented with inconsistent visual stimuli, that they can assess risks and that they can count up to three. These experiments have led some over-enthusiastic researchers to compare the intellectual capabilities of a hunting spider with that of a 1 year old human child with the obvious implication that, perhaps, these dear little creatures might be conscious.

Let us grant, for the sake of argument, that a hunting spider really does have the intellectual capability of a 1 year old child. What would this prove? To my mind it only proves that it is possible to program a neural network of 100,000 neurons to do some remarkable things.

Weight for weight, ants and bees have the largest brains in the animal kingdom. When two ants meet on the forest floor, touch each other's antennae and exchange pheromones, it is clear that they are communicating with each other – but what information is being passed is largely unknown. In the case of the honey bee, however, a remarkable discovery made by the Austrian scientist Karl von Frisch during the dark years of the Second World War proved that insects do, indeed, employ sophisticated means of communication to convey quantitative information about the direction, distance and quality of sources of food. Perhaps ants are doing the same but we have yet to discover the language they are using.

In recent years it has become fashionable to claim that crabs, lobsters, sharks and even goldfish are 'smarter than you think'. Apparently certain crabs can learn to negotiate their way through a maze and remember the route for up to three weeks and that lobsters can crawl out of 'tunnel traps' almost as easily as they can crawl into them; According to Culum Brown from Macquarie University, "Fish are more intelligent than they appear. In many areas, such as memory, their cognitive powers match or exceed those of 'higher' vertebrates including non-human primates."[20]

20 Brown, Culum (2004) *Animal minds: Not just a pretty face* New scientist, 2451: 42-43.

I am sure we have much to learn about the cognitive abilities of crustaceans and fish but the more sophisticated our own 'intelligent robots' become, the less need do I see to suppose that any of this behaviour is anything other than clever programming created by the same evolutionary processes that created eyes or digestive systems.

Artificial Intelligence

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 sort of 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 rule out the possibility that a computer made of wires and silicon will pass this test in the future. In fact the staggering advances which have been made in recent years in developing expert systems such as ChatGPT based on what are called 'large language models' using information culled from the world wide web would seem to suggest that the classic Turing test has at most a few years to run and that it will soon only be possible to distinguish the responses of a computer from those of a human being because the computer's responses are smarter! In fact, it would be wise to assume that it will be possible at some not too distant point in the future to create artificial beings which can emulate every aspect of our own human behaviour. Already such robots are being used as companions to the elderly and it will not be long, if it has not happened already, that a human being will fall in love with a robot.

But as I have repeatedly pointed out, my three indicative properties of consciousness – *empathy*, *imagination* and *intention* – should not be equated with *intelligence*. A super intelligent robot may be able to provide companionship to a lonely person, but will it ever derive any satisfaction from this work? A driverless car may soon be able to drive you to work – but can it be said to have *intended* to do this? Could it decide to go somewhere else? ChatGPT may be able to write passable poetry in a recognisable style – but could it invent a new style of poetry? A human being could well fall in love with an intelligent robot – but could a robot fall in love with a human being?

So we must be extremely cautious in ascribing consciousness to creatures who are amazingly intelligent. A recent television program has highlighted an astonishing dossier of apparently intelligent behaviours.

But the fact that a female firefly of one species can lure an unsuspecting male firefly of a different species to his death by pretending to be a member of his species does not imply that the so-called 'femme fatale' is indulging in conscious deception. Just because a shoal of fish can display what is called 'collective intelligence' in protecting themselves from predators does not mean that the shoal has some kind of conscious awareness that the individuals do not possess. And, just because mantis shrimps and dung beetles use the Moon and the stars to navigate their way back to their burrows, does not imply that they have taken a crash course in astronomy.

We must therefore be extremely wary of ascribing consciousness to any animal which displays merely intelligent behaviour. Equally we must be extremely wary of immediately ascribing consciousness to an animal which shows signs of empathy, imagination or intention as it is undoubtedly possible to program a mindless robot to display these qualities too.

On the other hand, when we do see an animal showing evidence of empathy, imagination or intention, we do not have to assume that they are mindlessly programmed that way because we know that, in at least one species of animal, Nature has used the phenomenon of consciousness to achieve these ends. (No human being has ever fallen in love, decided to retire or composed a symphony while in a coma.) So when chimpanzees show ample evidence of empathy, the fact that it may one day be possible to program a computer to simulate empathetic behaviour does not mean that chimpanzees are robots; just because it is possible to program a driverless car to remember a previously driven route does not mean that elephants are mindless automata; and just because it might, one day, be possible to program two robots to show affection for one another does not mean that swans must be unconscious.

If a creature shows evidence of *empathy, imagination* or *intention* the simplest explanation is still that it is conscious. Conversely, if the creature shows none of these traits, then, in all probability it is not conscious, not because it is impossible for an unconscious creature to do these things but because there is no particular advantage for such a creature to be conscious.

So what about octopuses?

Octopuses, together with squids and cuttlefish are often thought to show clear evidence of conscious behaviour. Octopuses will cover themselves with shells when crossing open ground in order to disguise themselves from predators (imagination?); cuttle fish in captivity have been known to take what appears to be a personal dislike to individual handlers who have mistreated them in the past (intention?) and an encounter with any of these creatures on a dive can often lead to an uncanny feeling that you are being watched by a creature who recognises you as another individual (empathy?).

The truth is, we do not know if cephalopods like these are capable of conscious thought or not. Unlike mammals and birds whose brains are basically the same as ours, an octopus has a completely different kind of brain. In fact, each of its eight arms is in some degree autonomous, controlled by a brain of its own. If an octopus is in any degree conscious, we can be sure that its thoughts are so totally unlike ours that it almost makes no sense to say that their subjective experiences are anything like our own.

There is another puzzle concerning cephalopods. All mammals and birds which clearly display the tell-tale signs of conscious thought, such as REM sleep and intention, generally live for 20 years or more. Consciousness has evolved in these animals in order for them to sustain long-term relationships with other members of the species. Cephalopods, however, rarely live longer than 5 years and live very solitary lives. In the rare cases where these creatures show intelligent behaviour in the form of apparent imagination or intention, it would, I think, be wise to consider first whether this behaviour could not possibly be instinctive before jumping to the conclusion that these creatures are conscious.

5 – Other Aspects of Consciousness

There are a number of other aspects of consciousness which I have not yet mentioned which may give us further clues as to which of our fellow creatures are conscious and what consciousness is for.

Self-Awareness

I mentioned earlier that being sentient – i.e. having a sense that there is an 'I' which is doing the sensing – is not quite the same as being *self-aware*. When a human baby first enters this world there is little doubt that it is in possession of a fully functioning sentient brain – but it cannot truly be said to be aware of its own *self* until several months have passed. It is therefore quite logical to suppose that there are other creatures which stay in this condition all their lives. Indeed, I suspect that a lot of people would be happy to put the majority of mammals and birds into this category. It is possible that to back up this proposition they would point to the MSR or mirror self recognition test.

If you habituate an animal such as a chimpanzee to the presence of a mirror in its enclosure and then, one day, while it is asleep, stick a piece of white paper to its forehead, then, as soon as it sees its image in the mirror, it will immediately touch its own forehead and remove the paper. Clearly the creature knows that the image in the mirror is not a real chimp but an image *of itself*. It could only do this if it had a concept of *itself* in the first place. Chimps therefore must be self-aware. They are not only conscious of sensations which come from outside; they are also conscious of their own bodies and the place of their consciousness within that body.

It would seem that the mirror self-recognition (MSR) test is a pretty foolproof way of detecting consciousness. The trouble is that relatively few animals pass the test unambiguously. The list includes the great apes and elephants but it definitely does not include dogs and cats. Magpies are said to pass the test but not parrots. The evidence for dolphins is not conclusive but recent experiments suggest that manta rays might qualify and there is strong evidence that the Cleaner Wrasse fish passes the test. If subsequent research confirms that some fish are indeed self-aware, this knocks my contention that consciousness is confined to mammals and birds into a cocked hat!

Now self-awareness is only a step away from the awareness of other creatures as individuals. It follows that if, as I contend, the primary purpose of consciousness is to enable creatures to recognise other creatures as individuals, then any creature which displays empathy must also be self-aware. The fact that dogs, for example, do not pass the test does not tell us that they are not self-aware; only that they haven't sussed out the physics of mirrors. Conversely, the fact that Cleaner Wrasse fish pass the test strongly indicates that we should be looking for empathy and possibly a social hierarchy within a shoal of that species. Fortunately for my theory, that is exactly what we find. Cleaner fish live in small groups in which a dominant male has control over a group of smaller females to the extent of 'punishing' those females who 'cheat' by eating the scales of the client fish rather than the less succulent parasites which it is the cleaner fish's job to remove. It may seem surprising to find consciousness in such a tiny creature but perhaps we should not be so surprised when we consider that the Wrasse's whole way of life is geared towards grooming other fish.

The upshot of all this is that self-awareness is not something which some conscious creatures possess and others do not; it is an essential part of what it is to be conscious. Sentience without self-awareness might be possible but is quite pointless.

Long-term memory

It is almost as difficult to understand how our memory works as it is to understand consciousness. 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 metaphors of a hologram or even that of a fractal algorithm which somehow enables you to reconstruct an image of a fern are probably more helpful. It should also be remembered also 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.

It is often said that 'elephants never forget'. I don't suppose that elephants are really any less likely to forget things than we humans; 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 other individual dolphins which they last met as long as two decades ago. Even rats, which have become familiar with several different mazes, can remember where the bait was placed last time they ran a particular maze for at least a week.

Nothing here suggests any necessary link with consciousness, though. Fish can learn new tricks and bees can famously memorise the position of food sources and communicate this information to other bees in the hive. On the other hand, there is an enormous gulf between the amount of information required to be stored by an elephant returning to an ancient water hole and the simple distance/direction information remembered by a bee returning to its hive after a successful foraging expedition. Likewise, there is a huge difference in the amount of learning involved between that of a sheep dog learning to round up sheep and that of a fish learning to press a button to obtain food.

Now it may be the case that consciousness and long-term memory are entirely distinct but I think it is significant that it is only those creatures which show evidence of empathy, imagination and/or intention that seem to possess long term memories. If, as I have suggested above, the main evolutionary advantage of consciousness is to enable the creature to engage in long-term personal relationships with other members of the same species then there is no point in being able to recognise your partner or other members of the tribe as individuals if you can't remember what those individuals looked like or what your relationship with them was like when you last saw them.

Conversely, I see little evolutionary advantage in being able to remember long distant places and events if all you are going to do is swim around a small area of coral or fly around a garden looking for food. Ants do not need to remember where their nest is located, they simply follow a trail of hormones back home like Hansel and Gretel in the fairy tale.

Emotion

Although empathy, imagination and intention are important indicators of the presence of a conscious mind, they are not the only identifiable characteristics of a conscious mind. Another is *emotion*. There is, however, no universally recognised definition of what constitutes an emotion or, indeed, how many different emotions there are. In the 1950's Paul Ekman suggested that there were six basic emotions: happiness, sadness, fear, disgust, anger and surprise.

Some of the things which we call emotions are not really emotions at all. The feeling you get in the pit of your stomach when you face a serious threat is not really an emotion but a pre-programmed response from your sympathetic nervous system which injects a cocktail of hormones into your blood stream, readying the body for a swift response. If we are to use the concept of an emotion to identify consciousness in other creatures we must be careful to distinguish the physiological aspects of the emotion from the psychological aspects. If you poke a stick into an ants nest the ants will show a behavioural response which it is easy to interpret as anger – they scurry around energetically and start biting anything which seems to threaten them. But is the ants nest 'angry'? I don't think so.

Since virtually all animals (including those like crabs and wasps which I maintain are not conscious) display the symptoms of anger and fear when put in appropriate situations, I do not think we should include these in our list of true emotions. In my view, an emotion is a purely *intellectual* response to a situation and may or may not be accompanied by a visceral response.

Removing fear and anger from Ekman's list leaves happiness, sadness, disgust and surprise. Do any animals show any evidence of any of these?

We are often guilty of using words which suggest that animals do, indeed, have emotional states. We say that hens are 'happier' when they have sawdust to peck at than if they are kept in cages with wire floors. It is certainly true that, given the choice, the hens prefer the former but this is not evidence that they are consciously happier – they may just be programmed to behave that way.

We might even describe courting grebes as being 'in love' but this is mere anthropomorphism. They are simply pair-bonding.

There is a bit more evidence that elephants can experience genuine sadness; it is said that young elephants will sometimes return to the site of the death of their mother and behave strangely and dogs can appear to grieve over the death of their master, but such stories are anecdotal at best and cannot really be said to provide definite evidence that animals can experience sadness.

Surprise is probably the most interesting one in the list. There seems to be no evolutionary benefit in being surprised by something unexpected but chimpanzees can be just as surprised by simple magic tricks as human children.

Later Ekman expanded his list of emotions to include amusement, contempt, contentment, embarrassment, excitement, guilt, pride, relief, satisfaction and shame. There is mounting evidence to show that chimpanzees exhibit symptoms of many if not all of these – from which we can pretty confidently conclude that chimpanzees are conscious. But this is not a big deal as few people doubt this contention today.

When it comes to identifying emotions (as I have defined them) in other creatures the problems of designing an experiment to identify the emotion in question are immense and the problems of interpreting the results even more so. If you find that a fish, having been accustomed to finding food in a certain place, appears to be surprised when the food is moved to a different place, are we justified in concluding that the fish is conscious? I doubt it.

Humans appear, in fact, to be by far the most emotional of all the animals. Why should I cry whenever I hear a particular piece of music? Why should Putin's invasion of Ukraine make me angry? Why do Paul Merton's jokes make me laugh? Why does the thought of people being tortured make me feel sick? It is certain that these emotions could only exist because I am conscious but that does not explain why they exist. It is also obvious that these emotions are only capable of being triggered because of the vast array of past experiences which form the background to my love of music, my sense of justice, my knowledge of the English language and my (fortunately limited) experience of pain

It seems to me that true emotions are not so much a characteristic feature of a conscious mind but an intellectual response to an event or stimulus and that the role of the conscious mind is simply that of a medium in which the emotion can be expressed. It follows that any creature which is capable of, say, being surprised must have considerable intellectual capacity and lots of long-term memories and that this is why it is only the most intellectual mammals which are capable of displaying any emotion.

Pleasure

Happiness is a state of mind – an emotion. As I have argued, only a few animals are capable of realising the intellectual response which constitutes an emotion. What I am referring to here is sensual pleasure – the pleasure of an orgasm or of being tickled, the pleasure of feeling the wind in your hair or of eating a delicious chocolate. Do animals feel this sort of pleasure?

Firstly, let us be clear that I am talking about a conscious subjective sensation here. Pleasure is the opposite of pain and in order for a creature to experience either, it *must* be conscious.

When we discussed whether animals feel pain, I pointed out that pain has an obvious function in helping to prevent the creature from doing themselves serious harm. But what is the purpose of pleasure?

I suppose the pleasure of an orgasm will encourage a creature to mate – but then what is the evolutionary benefit of masturbation? Why do primates (including us), dogs and cats love being tickled? Unlike other birds, crows positively revel in flying on windy days. Why do they do this? Penguins and parrots have been known to spend hours tobogganing down a snowy slope. How does this help them survive? Bears will (famously) go out of their way to rob a hive of its honey, but honey is not particularly nutritious so why put yourself in danger just for a few minutes sensual pleasure?

There are situations when the excessive indulgence of pleasure can be positively harmful to an organism, the most obvious example being drug use in humans, but fortunately there do not appear to be many cases of chimps tickling themselves to death or crows flying till they fall out of the sky! None the less, it is a puzzle as to why pleasure exists at all.

Many mammalian species learn a lot through play when they are young so this may be part of the answer but it cannot explain why so many adult mammals and birds seem to indulge in behaviour which does not appear to have any survival value, simply for the pleasure it apparently gives.

I suspect that the answer to this riddle is connected with the fact that, in my view, conscious creatures possess free will. Why do human beings cling so tenaciously to life? If life had no pleasure in it, the rational thing to do would be to commit suicide. Fortunately, for most of us, the pleasure we get out of eating a good meal, having sex, enjoying the warmth of the sun on our face etc. is sufficient reason to stay alive. Of course, dogs and cats, penguins and parrots do not realistically have the option of committing suicide and I doubt if the possibility ever occurs to them – but if pleasure had not evolved along with pain, our ancestors might have decided long ago that life was not worth the candle and we would not be here today.

The idea that the rational thing for a conscious creature with free will would be to commit suicide has occurred to many writers of science fiction. Think of Marvin the Paranoid Android in the Hitch-hiker's Guide to the Galaxy or HAL's descent into madness in 2001 a Space Odyssey. If you think about it rationally, there really is, literally, no point in living because, ultimately, there is no point in anything at all. You might argue that the purpose of life is to reproduce and pass on your genes; but what is the point of passing on your genes? You are not going to benefit. According to the most widely held view amongst scientists and philosophers, the world is governed by strictly deterministic laws[21] and has no purpose or goal – it just exists.

If this conclusion seems horrifying then I sympathise totally. Fortunately, however, conscious beings are bound to evolve with a sense of pleasure for the simple reason that if they evolved without this attribute they would, of course, commit suicide! There is one other comforting thought which we can deduce from this rather bizarre line of argument. If, at some time in the future, we are able to construct a supremely rational conscious robot, then, far from aspiring to kill all us humans and take over the world, its first (and last) rational act will be to pull out the plug!

Language

We have said very little so far about the role consciousness plays in permitting abstract thought. When a chimpanzee finds that the stick she is using is too short to extract a juicy grub from a tree, it is certain that she is not thinking (in English) 'this stick is too short, I must find a longer one.' But she is obviously thinking

21 See page 69

some sort of thought. Perhaps she thinks in pictures. Perhaps she sees in her mind's eye how a longer stick will do the trick and so she goes off in search of one. Having come to this conclusion, the only way she can teach her offspring what she has learned is by example. The range of grunts and whistles which she is capable of producing is not up to the job of communicating the idea that a short stick is no use, you need a long one.

The moral of this story is that, while consciousness permits logical thought (sapience), to communicate abstract ideas you need language and a voice box which is capable of producing a wide variety of different sounds.

Insects, crustaceans, cephalopods, and fish (including sharks) can only communicate with each other through colour changes and body language or by rubbing parts of their bodies together. Amphibians, reptiles and mammals have a voice box called a larynx which is capable of producing a variety of sounds. Birds have a rather different and unique organ called the syrinx which performs the same function. Although we are learning more and more about the sophisticated ways in which mammals and birds communicate, it remains the case that only *Homo sapiens*[22] has developed a way of expressing and communicating abstract ideas through language.

But once language has been invented something extraordinary happens.

Instead of just being a medium of communication between individuals, language becomes the medium through which individuals can indulge in abstract internal thought. This was a quantum leap forward that essentially resulted in *Homo sapiens* becoming the dominant species on this planet.

22 There is considerable debate about when in the evolutionary development of the human species abstract language appeared. It is currently fashionable to suppose that Neanderthals, whose brain and larynx was similar to our own, must have had language too but this is not proven.

Many volumes have been written arguing the extent to which language influences thought. This debate need not concern us here. The only point at issue is whether language and the kind of abstract internal thought processes which it allows could have evolved in an unconscious creature. I would argue that it could not for the simple reason that the creation of a language requires *imagination* – a feature which I have argued is exclusive to conscious creatures.

But there is an important caveat here. Language is so important to us humans that it is tempting to conclude that rational thought is *only* possible through language – but this would be a grave mistake. It is perfectly possible to hear a tune or see a picture in your mind and it is perfectly possible to compose a tune or imagine a design and write it down without language. Language may be unique to our species but rational thought is not. Indeed, it would also be a mistake to think that it is only *conscious* creatures which are capable of rational thought. When Deep Blue won a game of chess against the world chess champion Garry Kasparov in 1997, it is perverse to maintain that the machine was not 'thinking rationally'. If it was not thinking, what was it doing? OK – so you might want to reserve the verb 'to think' to apply only to conscious creatures, but then the question 'can machines think?' becomes meaningless. It is much more reasonable to define the phrase 'to think rationally' as the ability to process large quantities of data and come up with a solution to a problem which is consistent with the data. Under these terms, Deep Blue was certainly 'thinking rationally'.

Language has been crucial to the development of *Homo sapiens* and a language which is sufficiently rich as to include representations of abstract ideas could only be invented by a conscious creature, but, as I have repeatedly said, intelligence (i.e. rational thinking) is not the exclusive preserve of conscious beings.

6 – *Free Will*

Free will is such an important aspect of consciousness that it deserves a chapter of its own. The first question, however, is of course, does it exist? Do we really have the ability to choose our own actions or are we so constrained by our genes, our upbringing, our current desires and prejudices etc, that we have no choice but to act as we do? Do the laws of physics actually allow free will anyway?

If you were to ask 100 scientists and philosophers who have made it their business to study the question, I suspect that the vast majority would say that free will was an illusion and that it does not actually exist. If you were to ask the same question of 100 lay persons though I am sure that the majority response would be: “Of course free will exists – if it didn't people could not be held responsible for their actions. In any case, I just *know* it exists. I am as sure of the existence of free will as I am sure of the existence of my *self*, and for the same reasons.”

So what then are the arguments against free will that cause so many highly intelligent people to turn against what seems to be such an obvious conclusion?

Determinism

The main argument stems from the fundamental assumption that every effect has a cause. Over the centuries this (somewhat questionable) philosophical position has been crystallised into a scientific axiom called determinism which states that the future state of a physical system is completely determined by the state of the system at the present. Typically the behaviour of a system like the motions of the planets in the Solar System is governed by a small number of what are called differential equations. Given the initial conditions, these equations define precisely what state the system will be in immediately after, and immediately after that, and immediately after that *ad infinitum*. Even the motions of individual atoms which are governed by the equations of Quantum Mechanics (known as Schrödinger's equations) have this characteristic – they are completely deterministic (up to a point, that is).

If we accept this kind of determinism, the complete course of the universe was set in stone at the first instant of the Big Bang and nothing can change its future course. In a famous quotation the eighteenth century philosopher Pierre Simon Laplace said:

> *We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past could be present before its eyes.*

There is, obviously, no place for free will in such a universe. What will be will be and there is nothing anybody can do about it. Period.

This idea has caused a lot of controversy and a whole string of what are called 'compatibilist' philosophers from David Hume onwards have since tried to explain how free will, or at least the most important aspects of free will, are consistent with strict determinism. I have little time for their efforts. I can see how an *illusion* of free will can be compatible with determinism but that is not the sort of free will which I want to believe in.

In any case, there is one small flaw in the argument. I said that the laws of Quantum Mechanics were completely deterministic *up to a point*. Under some interpretations of Quantum Theory, a system evolves according to Schrödinger's equations but then at some point, nobody knows why, the system collapses at random and the equations have to be reset. A good example of this is that of a radioactive atom which may have existed for billions of years inside a lump of rock; then, completely out of the blue, it suddenly spits out an alpha particle. It is believed that this process is totally random. Quantum Theory can tell us precisely what the *probability* that the atom will decay in any given period of time is – but it cannot predict *exactly when* the atom will decay.

Now if the laws of physics permit genuinely random events (as I believe it does) then strict determinism is dead. But this does not really affect the argument against free will because random events at the atomic level are no better able to underpin the concept of free will than pre-determined ones. And even if there was some 'causal slack' in the behaviour of atoms, there is absolutely no evidence whatsoever that the behaviour of individual neurons is anything other than wholly deterministic. If there is any randomness at all in the behaviour of a neuron it will take the form of 'noise' which a healthy system will try to suppress. A brain which contains neurons which fire at random sounds more like a brain having some sort of fit, not a brain which is making rational choices.

In short, whether the laws of physics are deterministic or not, there is absolutely no place for free will. I think we must accept, therefore, that, under the laws of physics *as we currently understand them*, free will is impossible.

Now let's look as the arguments in favour of free will.

Our hypothetical lay person put forward two arguments in favour of free will, the second being the notion that each one of us just *knows* that we have it. Every second of our conscious existence we are somehow conscious of our *selves* and of the *decisions*, some great some small, which we we take in order to organise our lives. This subjective sense of being in control of our actions is just as strong as the subjective experience I have of sitting in front of a computer typing these words.

But if we accept that the laws of physics preclude free will then we must logically conclude that this subjective sense of being in control is just an illusion. Now it is well known that our subjective senses are easily fooled. Countless visual illusions exist which fool us into thinking that parallel lines are curved, that black is white and small things are bigger than large things etc. etc. And, of course, a brain under the influence of drugs can experience sights and sounds which do not actually exist. Notwithstanding these facts, it remains the case that for most of the time our senses can be relied on and the assumption that what we see and feel really does correspond to an objective reality out there is justified. So if we can (usually) rely on our five classic senses, why should we distrust our feeling that we have free will?

The objectors answer to this is simply: “I'm sorry. The laws of physics preclude free will so whatever you say, your sense of having free will is just an illusion. Live with it.”

What about our hypothetical lay person's first argument – that if free will did not exist then people could not be held responsible for their actions?

This argument is easily demolished too. If the world is deterministic then things just happen. When a judge orders a man convicted of rape to be put behind bars, he is not punishing the man for the deed, he is simply doing what the neurons in his brain, governed by the laws of physics, are telling him to do. When the King gives an OBE to a carer who has spent her life looking after orphaned children, he is not rewarding her – he is simply obeying the laws of physics. If, as I have said, in a deterministic universe every event which happens in the universe is set in stone from the beginning, then that includes not only the creation of our galaxy and planet Earth, the extinction of the dinosaurs and the Christmas tsunami of 2004, it also includes the conviction of the rapist and the award of the medal to the virtuous woman.

Even if the laws of physics contain some randomness, that only means that the course of events is not predictable, even in principle. It does not mean that the judge is making a moral judgment or that the good lady deserves reward. It just means that the random firing of neurons in the judge's and the king's brain happen to cause that particular result. If you think that this sounds extremely unlikely, I would agree with you; but if you truly believe in the laws of physics then you *must* conclude that *everything* that actually happens comes about either because it was inevitable from the start or because it was the result of some random event at an atomic level..

Another form of the argument from moral responsibility is this: "If we didn't have free will, then we would have no moral responsibility to curb our actions and we would all run amok and kill each other." This won't wash either. If the laws of physics preclude free will, then either we would all have run amok and killed each other long ago or the world would be exactly as it is. Since the former has not happened, we must conclude that the world is as it is because the laws of physics permit such a world. All this argument proves is that we *want* free will to exist, not that it actually exists.

But here's the rub. Yes, obviously the laws of physics *do* permit such a world – but how could such a world actually come about? Is it likely, or even conceivable, that a world in which criminals get punished and good people get rewarded is somehow an inevitable consequence of Schrödinger's equations? Let us look at this more closely.

The argument from evolution

I am perfectly prepared to accept that, given the laws of physics and the initial conditions at the Big Bang, the development of galaxies, stars and planets was inevitable. I am also prepared to believe that the origin of life on Earth was, if not inevitable, at least consistent with the laws of physics. I also am perfectly happy with the idea that life evolved through a process of Darwinian evolution to the point where there existed nervous systems which became conscious in some degree. As I have argued earlier, the main evolutionary benefit of a conscious brain was the ability to recognise other members of the same species as individuals and it was this ability (which I call *empathy*) which enabled some creatures to form strong pair-bonds and others to cooperate in hierarchical societies. I have also argued that the acquisition of consciousness brought with it two other abilities – the ability to create and invent new ways of doing things (*imagination*) and the ability to use long-term memories of past events to plan for the future (*intention*). Few species actually made much use of these last two abilities. Very few animals show much imagination and the extent to which sparrows and mice, even if they are conscious, plan for the future is seriously limited. Even *Homo sapiens* lived and hunted in small family or tribal groups for tens of thousands of years, inventing the occasional new tool or participating in ritual ceremonies etc. without seriously upsetting the course of Darwinian evolution.

But shortly after the retreat of the glaciers from northern Europe 12,000 years ago, something really dramatic happened. The invention of agriculture tied individuals to a particular piece of land; this led to the concept of ownership of property; people started to trade what they owned and record the transactions on clay tablets or papyrus leaves; trade enabled some individuals to forgo hunting for food and start specialising in making things such as clothes and tools for others to use; other individuals used their physical strength or intellectual advantages to dominate weaker individuals and put themselves forward as chiefs and kings or priests and religious authorities; still others, with time on their hands, turned to creating works of art and thinking about science and philosophy leading to the creation of things like the works of Shakespeare and the theories of Newton and Einstein; at the same time, religious leaders capitalised on primitive beliefs about the spirit world, developing different ideas about God and imposing different systems of morality on their followers. And so it came about that our current state of human society with all its faults and contradictions evolved.

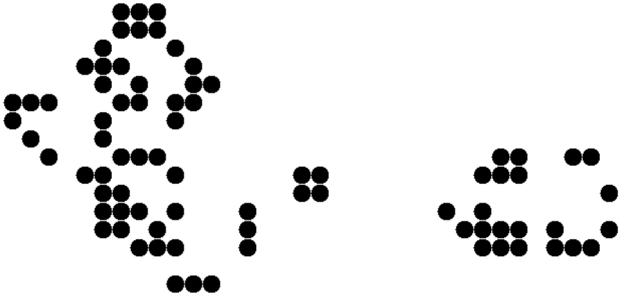
In my opinion, none of this creative activity could have come about without individuals making choices of their own free will. According to Darwin's theory, species automatically adapt themselves to changes in their environments and occasionally change so radically they turn into new species. But the corollary of this is that if there are no changes in the environment, a species which is in ecological balance with its environment will not change either. Early humans 50,000 years ago (and isolated groups like the Australian aborigines up to a few hundred years ago) were in ecological equilibrium with their environments and did not need to change. Obviously the retreat of the glaciers was in some way the trigger for the immense changes that came about in human society in the succeeding centuries, but Darwinian evolution cannot account for either the speed or the direction of those changes. In fact, the glaciers had retreated many times before but the only changes these episodes brought about was a northerly shift of the human population.

We cannot guess exactly why humans responded so differently this time round. I have suggested that it was the invention of agriculture but it may have been the invention of language or the development of writing. Experts will disagree. Whatever it was, it happened, and it did not happen through a process of Darwinian evolution. It was far too quick and far too dramatic a change. In my opinion, it came about because humans started using their conscious ability to imagine how things could be done differently and then to use their free will to do things differently.

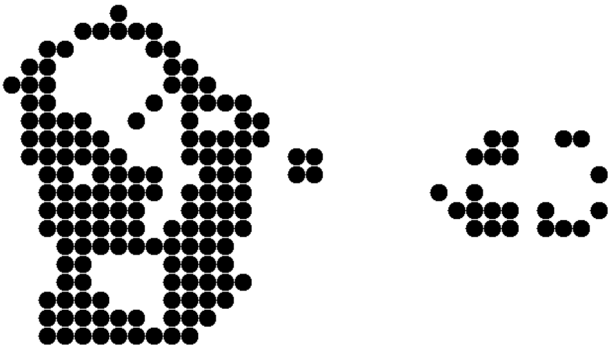
To see just how impossible it is to imagine how the current state of human society could have come about in the absence of free will, let me use a simple analogy.

The development of a deterministic universe can be simulated by Conway's game of LIFE. Without going into details the game is played on an infinite square grid each of whose cells can be either black or white. Fixed rules determine how any given configuration evolves into a new configuration. It turns out that there is no way the future evolution of a given configuration can be predicted in advance – the only way to find out what is going to happen is to run the game and see. Some initial configurations live for a while and then die, but a few appear to grow without limit. Here, for example is a five spot configuration that grows for a while at least:

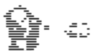
After 60 generations this develops into:



Obviously as time proceeds the number of possible configurations into which it could develop increases rapidly. Conversely, as time proceeds the probability that an initial configuration hits a specific recognizable target decreases in proportion. What then are the chances that an initial configuration, similar to but different from the pentomino illustrated could generate the following picture?



t

(In case you do not recognise the image, here it is much reduced:  . It is a digitised copy of a famous image of Marilyn Monroe by Stefano Padoan.)

As it happens, the chances of this configuration arising are zero because *there is no antecedent* which could produce this exact configuration. This is a consequence of the rules of LIFE being entirely deterministic. If the rules contained some randomness, then this precise configuration could result by chance but even with this simple example, if we ran the game on the fastest computer in the world it would take longer than the current age of the universe to generate even this simple image[23].

If we translate these results into the context of the history of the universe, it is either astronomically unlikely or even impossible for there to exist a configuration of the early universe which could develop into a world which contains the works of Shakespeare or the Newtonian theory of gravity. The only way in which sophisticated societies could develop in which the sciences and the arts could flourish and in which systems of legal and moral responsibility could develop is a universe in which conscious creatures have free will.

The argument from creativity

The crux of this argument is the idea that the process of creativity, whether it is the invention of agriculture. construction of a moral code or the writing of a symphony, absolutely requires a conscious brain capable of free will.

I mentioned on page 43 a pod of orcas which has invented a new way of catching seals. I cannot believe that four autonomous robot submarines, possibly equipped with a random number generator, pre-programmed with the goal of killing seals could come up with the idea of coordinating their actions in such a way as to wash the seal of his ice floe. It is just not going to happen – not in a billion trillion years anyway. They are just going to go on killing seals in the way they have been programmed to do.

The whales, however, equipped as they are with conscious

23 The image contains about 400 pixels each of which can be either black or white. The total number of different possible images is therefore 2^{400} which is equal to about 10^{120}. If a computer could generate a million images per second, it would only have checked about 10^{23} images in 5 billion years so my claim is a gross understatement to say the least!

brains capable of imagining things which have never been imagined before, have seen a possibility and made a conscious decision to do something different.

When Shakespeare was writing a play or Mozart was composing a symphony, they were both continually making conscious decisions about writing this word or composing that note using their powers of free will – and without free will they could never have created these fabulous works of art.

When talking about free will we often try to simplify things to make our ideas as clear as possible. In discussing free will we tend to concentrate on binary decisions like whether or not to accept a job or whether to choose chocolate or vanilla ice cream. Mark Balaguer calls these 'torn decisions'[24] and suggests that we may only make such decisions a few times a day. I think he underestimates the importance of free will. I think that we are making 'torn decisions' every moment of our waking day; whether is is to say 'hello' rather than 'good morning' or to type 'which' instead of 'that'; whether to scratch an itch or to rub it, whether to think about a problem at work or forget it and go to sleep.

So now we have two slam-dunk arguments. The laws of physics as we currently understand them preclude free will but if we didn't have free will we would still be living in the Stone Age and the works of Mozart and Shakespeare, Einstein and Newton would not exist.

For the moment I will leave you to guess where this is going.

24 Mark Balaguer *Free Will*

7 – Theories of the Mind

So far we have discussed various 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.

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 which has sentience and (possibly) free will? Is it just a question of organization? Is consciousness an inevitable consequence of the way the neurons in the brain are organised or are we missing something completely? Are we going to have to take paranormal phenomena like telepathy and extra-sensory perception seriously in our search for the key to consciousness?

The great majority of neuroscientists these days would incline to the belief that there is nothing more to the conscious brain than what we already see and know about. It is just a vast collection of neurons wired up together in an incredibly complex manner and out of all this complexity, consciousness emerges naturally in the same way that the thermodynamic properties of a gas emerge from the mechanical behaviour of its individual molecules, and the properties of a living organism are simply a consequence of its incredibly complex chemistry. This belief (and it is a belief, not a fact) is known as the Computational Theory of the Mind and deserves close scrutiny.

The Computational Theory of the Mind

In the last 50 years, digital computers have developed from room-sized behemoths, which took a week to test a 3000 digit prime, to hand held devices which can triangulate a GPS position to within an accuracy of a couple of metres in a few hundredths of a second. In the same half century the assumption that the human brain is just a fabulously complex computer has become so entrenched in popular culture that it is almost heretical to consider the alternatives. And this is not without good reason. Nothing we have discovered about the detailed behaviour of the 86 billion neurons that we presume are responsible for whatever thinking goes on inside our brains shows any sign that they do not obey the classical laws of physics; in fact, each neuron appears to be just a straightforward logic circuit with inputs and outputs, just like a digital one. (Admittedly a typical neuron may have several thousand inputs and outputs and it uses a multi-level coding system not just a binary one but there seems to be no reason in principle why the action of any individual neuron could not be mimicked precisely with a small microprocessor chip.) The implication of this is that if you could replace every neuron in my brain with a suitable integrated circuit[25], the resulting machine would do exactly what my brain does. It would *be* a conscious brain and, presumably it would *have* a mind.

Now this is not just philosophical speculation. This is a proper scientific theory with profound implications and testable consequences and it imposes some impressive limits on what computational brains can achieve. For this reason (and others) it

25 The diameter of a neuron in the human brain is about 10^{-6} m. The thinnest wires we could use to construct an artificial brain are about 10^{-4} m in diameter. It follows that a brain made from copper wires and silicon chips would have to be at least 100 times bigger than a human brain and would therefore be the size of a block of flats. Given the necessity of supplying power and cooling to all the chips, a more realistic estimate would be a building the size of the Pentagon!

has come in for a lot of serious criticism. Let us have a look at some of the implications of the theory and how these implications can be used to attack the theory, together with some of the counter arguments.

Objection No 1. The Computational Theory of Mind implies that the human brain is only capable of carrying out tasks which can be performed by a Turing Machine[26]*. Roger Penrose has argued on the basis of Gödel's famous theorem*[27] *that since human brains are capable of proving certain mathematical truths which it is impossible for a Turing Machine to prove, human brains cannot be Turing Machines*[28].

If this argument is upheld, it would prove beyond doubt that the human brain is at least in some respect non-computational but although Gödel's theorem is accepted by the mathematical community, its relevance to the issue of the human brain is widely disputed.

Objection No 2. A Turing Machine cannot create anything; it cannot even, for example, print out a random number let alone write a novel or compose a symphony. It follows that a computational brain could not do these things either.

The problem with this is that it is impossible to define what we mean by a random number let alone a 'novel' or 'symphony'. In any case, it would be easy to equip a standard Turing Machine with a random number generator and in the course of time I have little doubt that it *will* become possible, if we should so desire, to program a computer to write simple novels and compose

---

26 A Turing Machine is an ideal computer which can be programmed to carry out any logical operation whatsoever including proving the truth or falsity of any mathematical statement.

27 In 1931 Kurt Gödel proved that there existed certain mathematical statements which could be shown to be true but which could not be proved or disproved by a Turing Machine

28 Roger Penrose: *The Emperor's New Mind*

acceptable music by using a random number generator and a huge bank of literary and musical phrases.

Objection No 3. A Turing Machine equipped with a random number generator might be able to write a hack novel or a halfway acceptable symphony but it could not invent a completely new plot or compose something in a completely new style.

Now we are getting somewhere. Human brains do seem to be remarkably inventive (but not uniquely so). However, it is by no means obvious that a Turing Machine with a random number generator is logically incapable of composing something completely new. And in any case, even if you propose that the human brain is, in some way, non-computational, you still have the problem of explaining how it can create something that did not exists in any shape or form before it was created.

Objection No 4. A Turing Machine can solve logical problems but it doesn't understand what it is doing.

This is the essence of John Searle's objection to the computational theory of mind exemplified by his famous 'Chinese Room' thought experiment in which a man, locked in a room but with access to a huge database of instructions (in English) as to how to respond to questions posed in Chinese, can apparently answer questions in that language without understanding a word of it. The trouble with this argument is that we have now moved too far away from what is scientifically testable because the concept of 'understanding' is not sufficiently well defined. If we define the concept, for example, as 'the ability to respond appropriately to complex queries by drawing together many different aspects of knowledge and data from many different sources' then it is clear that, while the individual in the room does not understand Chinese, the *system* (which comprises the subject plus the dictionaries and data bases which the room contains) does.

> *Objection No 5. A Turing Machine cannot experience anything. A Turing Machine equipped with a camera could distinguish red from green but it cannot experience redness or greenness*

Things which we can experience like redness and greenness are called qualia and their existence has been hotly (and in my opinion, fruitlessly) debated over the centuries by philosophers. Yes, I experience qualia but it is quite pointless to attempt to prove that other creatures and machines do or do not because the whole point about experiences is that they are unique to the individual who experiences them. They are completely subjective and hence outside the remit of objective scientific enquiry. So, while it may seem outlandish to suggest it, it is perfectly consistent for the computational mind theorist to maintain that a television camera pointing at a red rag is 'experiencing' something analogous to the sensation we have when looking at the same object. The only difference between the two 'experiences' is in the amount of processing which goes on in interpreting the scene. According to the computational theorist the camera does little more than signal the colour of the object; the human eye-brain combination, however, will instantly interpret the scene as a communist flag, or a matador's cloak or whatever. But this difference is only a matter of degree. Sentience, it is argued, is just very sophisticated information processing.

> *Objection No 6. A Turing Machine cannot feel emotions, it cannot be conscious and it cannot make decisions of its own free will.*

Now we are so far away from objective scientific enquiry as to make the objection meaningless. Emotions like love and anger are experiences which don't even have recognizable stimuli so how are we to tell whether or not a Turing Machine is capable of feeling these things? Maybe consciousness is an automatic by-product of any sufficiently complex information processing. And as for free will, who is to say that we have it anyway? In fact all the computational theorist has to say in the face of these objections is that, while there is no objective proof that a Turing Machine is incapable of feeling emotions, being conscious and exhibiting free will, ample proof of the contrary proposition is plain for all to see, sitting on your own shoulders!

So where does this leave us?

So far all we have shown is that the computational theorist has sound counter arguments to all the objections which have been raised to the idea; but this does not prove that the idea is correct, only that it is logically possible. I am still left with the nagging feeling that we have missed something somewhere. I am quite prepared to accept that the brain of a nematode worm, consisting as it does of just 302 functional neurons, is entirely hard wired; and it seems perfectly feasible to me to suppose that all *unconscious* creatures from insects to sleeping humans have computational brains too; but surely there is a quantum leap from an unconscious brain (e.g. a human in a coma) to a conscious one. How are we to explain the huge change in behaviour when a conscious creature is rendered unconscious by means of an anaesthetic drug or when they fall asleep? Why is it that anaesthetic drugs affect only consciousness? If the effect of such drugs is to disrupt the ability of neurons to carry out their job correctly, why do they only affect those neurons responsible for consciousness? If, as a recent author[29] has suggested, consciousness is brought about by some sort of coordinated global activity in the brain, why is it that it is only this activity which is shut down when we are drugged or fall asleep?

I do not believe that it is sufficient for the computational

29 Stanislas Dehaene: *Consciousness and the Brain* p164

theorist just to say that consciousness is an inevitable emergent property of a brain which is wired up in the correct way. If that was all there was to the whole phenomenon of consciousness, then why would any creature, blessed with all the assumed advantages that consciousness confers, submit to being rendered unconscious by a whiff of chloroform, or to voluntarily give up these advantages for 8 hours every day?

So if consciousness is not an *emergent* phenomenon, what kind of phenomenon is it? Are there any precedents in the history of science where we have encountered phenomena which have appeared to be emergent but have subsequently turned out to be something else?

First we need to be very clear what we mean by an emergent phenomenon.

Emergent properties

Emergence is all the rage these days. It has been used to explain everything from the weather to the origin of life. But emergence appears to be a hard concept to define and a cursory survey of books and the internet will reveal many different definitions. Broadly speaking, however, they all boil down to the observation that many systems that are made of large numbers of smaller entities seem to have properties that the smaller entities do not themselves possess. In short, the whole is often more than the sum of its parts.[30] For example:

- The behaviour of an ideal gas can simply be deduced from the behaviour of atoms and molecules moving under the influence of Newton's laws.
- The movements of a shoal of fish or a mumuration of

30 For a comprehensive list of definitions see: http://www.vub.ac.be/CLEA/dissemination/groups-archive/vzw_worldviews/publications/wvdefemerg.pdf)

starlings can be easily simulated by a computer program obeying some simple rules such as follow your leader but don't get too close to your neighbours.

- Even if its precise strength and course may be unpredictable in practice, the properties of a hurricane can be reduced to and deduced from the properties of masses of air of which it is formed.

All the above are examples of what David Chalmers calls *weak emergence*[31]. The behaviour of the system may be unexpected but it is not difficult to see how the behaviour comes about as a result of the interaction of all the component parts.

Sometimes, however, the connection is not so easy to see. For example, a really complex computer program like the chess program Deep Blue or a large neural network can produce behaviour which was unforeseen by the programmer (and may even be, in some sense, impossible to predict *even in principle*); the properties of a chaotic system like the rings of Saturn or the emergence of structures like the 'gliders' in Conway's game of LIFE were never predicted before they were observed and there is no way they could have been predicted in detail because there are no general laws saying that if a chaotic system is in a state X now then it will approximate to state Y after a given interval of time. The only way, in fact, to predict the state of a chaotic system in the future is by running a sufficiently accurate simulation of it.

All these examples suggest a powerful means of identifying emergence. If you can simulate it on a computer, then you are dealing with (weak) emergence. But what if you can't even simulate it on a computer? Take, for example, the economic law of supply and demand which asserts that if supply exceeds demand, prices will fall and vice versa. Just as the gas laws follow from a consideration of the behaviour of molecules, this law

31 David J. Chalmers: http://www.consc.net/papers/emergence.pdf

follows from certain assumptions about the behaviour of individuals buying and selling; but unlike molecules which obey well known physical laws, we cannot reduce the behaviour of individuals to simple physical laws. It is not that individuals behave in ways that disobey the laws of physics, it is just that the gap between the laws of physics and the phenomenon in question is just too vast to bridge.

Or what about the rule of insular dwarfism – the tendency for large species confined to a small region to evolve into smaller species. There is no way in which you could run a computer simulation to predict this law; and yet there is little doubt that it comes about as a direct result of the Darwinian laws which govern the development of individuals within a species.

It is tempting to regard examples such as these as examples of what Chalmers calls *strong emergence*, but he is adamant that there is only one example of strong emergence and that is – *consciousness*. I think it is a great pity that Chalmers chose to define strong emergence so restrictively. There is a strong case for regarding the two examples given above as being qualitatively different from the examples of weak emergence. Chalmers defines strong emergence as phenomena which are not deducible *even in principle* from the low level laws from which the behaviour emerges.

Now we can argue at length over what he meant by the phrase *even in principle* but it is clear from what he says later is the essay that he regards strong emergence as necessitating fundamental changes to the laws of physics:

> *... if there are phenomena whose existence is not deducible from the facts ..., then this suggests that new fundamental laws of nature are needed to explain these phenomena.*

Now regardless of whether the laws of supply and demand or insular dwarfism are regarded as examples of strong or weak emergence, nobody would go so far as to suggest that we need new laws of physics to explain either. If there are any genuine cases of systems which show behaviour which transcends the laws of physics then they should be dignified with a grander name than merely strongly emergent. I propose to call them *transcendent* phenomena. I happen to agree with Chalmers that consciousness is one of them but are there any other examples of transcendent phenomena? Unlike Chalmers, I believe that there are. An obvious historic one is the Second Law of Thermodynamics which states, among other things, that the entropy (or disorder) of a closed system will always ultimately increase.

I have already mentioned that it is possible to deduce the gas laws by applying Newton's laws to the motions of individual gas molecules and throughout the nineteenth century (and even to this day) great efforts were made by Boltzmann and others to derive the second law in a similar way. All attempts have, however, failed. The Second Law of Thermodynamics simply cannot be derived from Newton's Laws of Motion. You can easily see this for yourself by running a video of a snooker break in reverse: what you see will obey Newton's laws precisely – but it will dramatically violate the second law of thermodynamics.

In order to explain the second law, Newton's laws of motion must be modified. The simplest way to do this is to assume that there is a tiny bit of randomness whenever two gas molecules (or two snooker balls) collide[32]. In the nineteenth century the second law was therefore a prime example of a transcendent phenomenon requiring new physical laws to explain it. Now we can regard it as merely weakly emergent.

Are there any current unexplained phenomena which may

32 I should point out that this interpretation or explanation of the Second Law is the author's own and is not universally accepted.

require a revision of any of our fundamental laws?

Superconductivity[33] was originally thought to be restricted to materials very close to absolute zero in temperature but in 1986 Bednorz and Muller discovered the first material which is superconducting at 30K (30 degrees above absolute zero) and soon after, several materials were discovered that superconduct at the temperature of liquid nitrogen (77K).

No completely satisfactory explanation of this phenomenon exists. Most scientists believe that this is simply because the problem has (so far) proved to be extremely difficult and that, like all other properties of matter, the phenomenon is emergent and that eventually an explanation will be found. It may, however, turn out to be the case that the laws of physics as we currently understand them are inadequate to explain the phenomenon. In which case the phenomenon transcends the currently known laws of physics.

And what about some of the really big issues? Is the origin of life emergent or transcendent? Are the laws of physics adequate to explain how life evolved or not?

There was a time when life itself was regarded as a transcendent phenomenon which depended on a force or substance which was outside the laws of physics – an attitude known as vitalism. We now know enough about the chemistry of life to reject this position but the origin of life remains a mystery.

And finally: is consciousness merely an inevitable consequence of the way 100 billion neurons are wired together or is it a completely new phenomenon which transcends the laws of physics?

We have already noted that almost all scientists who study the

33 Superconductivity is a state in which a material has zero electrical resistance.

conscious brain tacitly assume that consciousness is an emergent phenomenon wholly reducible to the workings of neurons and other cells in the brain. But what if they are wrong? What if there genuinely is something about the conscious brain which makes it qualitatively different from the unconscious brain and which enables it to transcend the laws of physics *as they are currently understood*. Are we really so arrogant as to assume that we now know everything there is to know about the way the world works? Just consider the following short list of questions we currently cannot answer:

- why is there so little antimatter in the universe?
- what is dark matter?
- what is dark energy?
- do gravitons exist?
- what enables high temperature superconductivity?
- how can two separated photons be entangled together?
- How can Einstein's General Theory of Relativity be reconciled with Quantum Theory?

Faced with a list of questions such as these, it seems premature to claim that we know everything there is to know about how the brain works – even in principle. Surely it is more rational to keep an open mind and hold fire on publishing books with titles like 'How the Mind Works'[34], 'I am a Strangle Loop'[35] and 'Consciousness Explained'[36]. The authors of these books all have their own pet phrases with which they, frankly, bludgeon the reader into believing that they explain how consciousness arises in the human brain. None of the authors, however, attempt to provide any evidence for their theories or consider the extent to which other creatures might be conscious. The word 'sleep' does not even occur in the index of any of them and free will is either barely

34 Steven Pinker
35 Douglas Hofstadter
36 Daniel C. Dennett

mentioned or explicitly denied!

One of the most popular current theories of the mind is the 'global workspace' theory[37]. The argument goes something like this: The human brain is a massively parallel computer in which different parts of the brain are carrying out different functions like monitoring breathing, interpreting visual sensations etc. etc. largely independent of one another; but during conscious thought the brain becomes a single interconnected neural network with multiple feedback loops; neurons which connect widely separated areas of the brain start to share information on a global scale, and it is this global sharing of information which is somehow responsible for the phenomenon of consciousness.

My objection to this plausible picture is that I just do not see how any neural network which is bound totally by the currently known laws of physics could ever possess a sense of self and therefore *empathise* with another similar brain; or how it could *invent* a genuinely new idea or, above all, how it could freely choose what it *intended* to do next – in other words, to exercise its *free will*. Let's start with imagination.

Imagination

Are there really things which a human brain can do that a classical, computational computer cannot do? Anyone who has played with the latest generation of ChatBots must surely have been impressed by their capabilities and there are genuine and well-founded concerns that Artificial Intelligence will impact our lives in the not too distant future in ways that it is impossible to foresee. It may sound like science fiction but the day may not be far off when all our clothes and cars are designed by computer; when our pop songs and film scores are written by computer; when our newspapers and books are edited and authored by computers and when computer-generated art and music finds itself

37 See "Consciousness and the Brain" by Stanislas Dehaene

in prestigious art galleries and concert halls.

So what is left for humans to achieve in the way of creation?

Since the dawn of civilization, mankind has demonstrated its creativity in five major areas: Art, Music, Literature, Science and Mathematics. It looks as if computers are on the verge of hijacking the first three. What about Science? Could a classical computer have come up with Kekulé's idea that the six carbon atoms in benzene were joined in a ring? Could a classical computer have come up with Crick and Watson's double helix model of DNA? Will a computer soon tell us what dark energy really is? Can we expect a classical computer to explain to us the mysteries of Quantum Theory? Somehow, I doubt it. But that does not prove that it could not happen. As I have already said, if, as the majority of neuroscientists believe, the human brain is a classical, computational computer, then we have all the proof we need that there is no logical impediment to a classical computer doing all these things.

What about Mathematics?

Here we are on rather more solid logical ground. Suppose, for example, you ask a Turing machine equipped with the simple rules of arithmetic to find two numbers which add together to make 10 and multiply together to make 40, the machine will (correctly) reply that two such numbers do not exist – and this is exactly what ChatGPT3 concluded when I posed this very question. However, in 1545, Girolamo Cardano put forward a solution to the problem. He claimed that the two numbers were $5 + \sqrt{-15}$ and $5 - \sqrt{-15}$. Of course, Cardano knew that it was impossible to take the square root of a negative number and it was this fact that caused ChatGPT3 to say that there were no solutions. But Cardano's non-computational brain was able to step outside the box and argue as follows: suppose that the square root of minus fifteen really does exist then when you add these two numbers together, the square root terms cancel each other leaving you with 5 + 5 = 10. Multiplying them together is only a little more difficult. The arithmetic goes as follows:

$$(5 + \sqrt{-15}) \times (5 - \sqrt{-15}) =$$
$$(5 + \sqrt{-15}) \times 5 - (5 + \sqrt{-15}) \times \sqrt{-15} =$$
$$5 \times 5 + \sqrt{-15} \times 5 - 5 \times \sqrt{-15} - \sqrt{-15} \times \sqrt{-15}$$

The two middle terms cancel out and, of course, by definition $\sqrt{-15} \times \sqrt{-15} = -15$ so we have

$$(5 + \sqrt{-15}) \times (5 - \sqrt{-15}) = 5 \times 5 - -15 = 25 + 15 = 40$$

The idea that numbers like $\sqrt{-15}$ really do exist led to a whole new branch of mathematics, the mathematics of complex numbers, without which the modern world would be literally impossible to imagine.[38]

Other similar examples include Lobachevsky's discovery of non-Euclidean geometry or the development of wholly new branches of mathematics such as topology or calculus. I do not believe that any of these ideas could have come from a computational machine however sophisticated its programming because, before the idea was invented, the program could not contain the idea, even implicitly. The only way that a computational machine can create something really new is if it is equipped with some sort of random idea generator. Solutions or ideas would be generated at random and the machine would simply use its vast stores of data to eliminate incorrect or useless ones.

This is how evolution works. Random mutations are selected

38 When ChatGPT3 was given the problem of finding two *complex* numbers whose sum was 10 and whose product was 40 it correctly set up the four relevant equations and should have come up with the correct answer. Unfortunately it made a pig's dinner of simplifying the equations and ended up making an elementary algebraic error which caused it to come up with the numbers 41 and -31! Hopefully this bug will have been corrected in subsequent version of the ChatBot!

for their beneficial effects; adverse mutations are ruthlessly eliminated. But evolution works over hundreds of generations and millions of years. Human inspiration is not random and works much faster. Did Einstein consider all the possible letters of the alphabet before coming up with $E = mc^2$? Not at all. It was simply the inevitable consequence of 'thinking outside of the box'.

So if the human brain is not a classical computational computer, what is it? Three possibilities spring to mind: a) it is a vast neural network with a random number generator; b) it is a quantum computer; c) it works by magic.

What are the arguments against its being a vastly complex neural network out of which consciousness emerges when its feedback loops become sufficiently tangled (or whatever)?

The first thing to say is that it is mathematically proven that any neural network can be functionally simulated by a Turing Machine – so all this talk about vastly complex neural networks and multiple feedback loops etc. is a complete smokescreen. Fundamentally, we are still talking about a Turing Machine whose sole function is to translate a string of 0's and 1's on a piece of tape and output another string of 0's and 1's. Each time it has the opportunity to erase the digit on the tape and write a new one, it does it either in obedience to the instructions which are coded in the current state of the machine or it uses its random number generator to print a random digit. I can see no suggestion of imagination or creativity coming into this process and even if it were to be repeated a billion trillion zillion times, I still can't see the machine coming up with the idea of complex numbers or the principle of Relativity except by complete accident. Nor can I see how a computational machine can exercise what I understand to be free will. Either the results of its deliberations are completely determined by its initial state, or they are, in part, random. Either way, the machine is incapable of making a free, reasoned choice between alternative courses of action.

In my view, therefore, the idea that the conscious brain is a computational machine is simply ridiculous. Consciousness cannot be weakly emergent; it must be a *transcendent* phenomenon. And our best guess as to where that new phenomenon is to be found is in quantum theory.

Is there anything at all in our current understanding of quantum theory that might break the logic of the above argument and suggest a way in which a quantum brain might be able to generate genuinely new ideas? Sadly, I have to concede that the answer to this is, currently, no. But I live in hope. The main property which the designers of the current generation of quantum computers claim of their creations is that they can perform calculations much faster than classical computers because they can, in a sense, perform several calculations simultaneously; the rationale for this being that the fundamental unit of calculation (the 'Qubit') can hold several different values at the same time. Now there are almost as many interpretations of quantum theory as there are quantum theorists but one of the most popular is the 'Many Worlds' interpretation which can be construed as implying that a quantum computer does its simultaneous calculations in a series of parallel universes. This may sound incredible but the quantum computing industry is currently worth $812 million in 2022 and is projected to be worth nearly $10 billion by 2023 so there are some very wealthy companies prepared to throw serious money at the idea.

For myself, having swallowed the idea that there could exist an infinite number of parallel universes out there, I no longer find it particularly hard to believe that if a quantum brain had access to all these universes, an idea which would take millions of years to occur to a Turing Machine might not occasionally present itself to a receptive brain in a flash of inspiration, as it were.

How about selfhood and free will? Does the idea that the conscious brain is a quantum computer help us to understand or explain either of these concepts?

Selfhood

When I go into non-REM sleep I, apparently, forget everything I ever knew. My body becomes a mindless robot. It breathes and digests my supper and it can respond to loud or violent stimuli but it can't solve a problem, it can't invent a new device or decide where to go next on holiday. But when I wake up the next morning, it seems as if I can seamlessly take over where I left off the previous night in solving my problem, developing my ideas for a new device or widening my holiday researches. In short, it seems obvious that the 'I' (the self) which went to sleep is the same 'I' (or self) as the one that wakes up the next morning. But this is a mistake. For it to be true, the self would have to have some sort of independent existence of the body to which it is usually attached. While this notion is appealing to many and has been the default assumption of the majority of the world's religions, the concept of an independent self is full of paradoxes. What actually happens to the self when the subject goes to sleep? What happens when the subject dies? What if it became possible to teleport a human being molecule by molecule to a distant planet? How would the self know which planet to go to? What if it became possible to clone a human being molecule by molecule? To which clone would the self get attached?

If we are to develop a proper scientific understanding of consciousness we must reject the idea that selves can exist outside a conscious brain. If we wish to keep using the word, we must define it carefully.

I think we can agree that the most important things which remain the same after a period of sleep, and which essentially define the self we are talking about, are the memories which the subject's brain contains. If a subject suffers accidental trauma to the brain which significantly affects their memories or habitual behaviour (which is a form of memory) then we are justified in saying that the person is not the same self as he was. Likewise the gradual loss of memory associated with dementia or Alzheimer's disease can be described as a gradual loss of selfhood.

One important corollary of this idea is that people with no memories at all cannot really be said to possess a self or even be a person. The most obvious example is that of a new born baby. A foetus develops a fully formed brain after around 32 weeks gestation and during the remaining 8 weeks before birth it shows typical cycles of REM/non-REM sleep so we can presume that it is dreaming and building up a bank of memories and that it is conscious some of the time. But at the moment of birth, of what can the baby be conscious when she first opens her eyes? With no relevant memories with which to compare the avalanche of sensations which flood her brain, she cannot be expected to be able to interpret them or sift out the important ones. It will take several weeks before she can differentiate her mother's face from a stranger's, or even a red light from a blue one. A new born baby, therefore, does not have a sense of self. How can it be aware of itself (or its self) when it cannot even be consciously aware of a bright light or a loud sound?

In short, when I refer to my *self* what I am really referring to is the sum total of all my past experiences which are held consciously or unconsciously in my brain. Advocates of the computational theory of the mind should be quite happy with that definition. You could say that the 'self' of a computer was simply the sum total of all the information stored on its hard disc. If you were to take the hard disc out of one computer and put it into another, the second one effectively becomes the first; likewise if you copy the hard disc you now have two identical computers.

But of course, what this analysis leaves out is any explanation of *sense of self*; any sense of a subjective 'I'; any feeling of *sentience* or of *being*. For the computational theorist, this subjective sense of self is simply an emergent phenomenon consequent on the massive interconnectedness of the neural network of the brain and of the way it is organised. But if we admit that consciousness is consequent on the existence of past experiences, then we must give those memories a role in explaining the subjective nature of our sense of self.

Now we must admit that we haven't got the first clue how memories are stored in the brain. How is it that we can recognize a face in a photograph that we have not seen for years? How is it that we can remember the meaning of an obscure word which we have not come across for ages? How is it that the sound of a bell can resurrect distant memories of school? All these things are stored in our brains and make up who we are and they are all potentially available to us (though often difficult to access) whenever we are conscious. Somehow, massive interconnectedness of neurons does not seem to me to cut the mustard. Even with 100 billion neurons, there does not appear to be enough information capacity in the brain to store everything which I can remember. Quantum computers seem to be on the verge of enabling us to do things which conventional computers cannot do. Is it too fanciful to suppose that there is a mechanism at work in the conscious brain which enables our brains to do things which a computational brain could not do? Could this mechanism enable the brain to store long-term memories more like the way a hologram stores the image of a chess piece or a fractal algorithm stores the image of a fern rather than the bit-by-bit way that a computer stores a photograph? And is it beyond imagining that this holistic brain which can communicate internally via methods of which we currently have no understanding could also give rise to that subjective sense of self which we all experience so intensely?

Without pushing the idea too hard, is it not conceivable that a Bose-Einstein condensate, that is to say a large collection of atoms, cooled to near absolute zero and which behave collectively as a single quantum object, might be on the verge of possessing a sense of self?

OK – the idea is ridiculous; but the idea that the Earth moved or that it might be possible to speak to someone on the Moon was once thought ridiculous so it might be wise to keep an open mind.

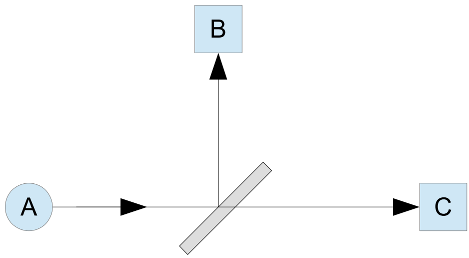
Free will

My final objection to the computational theory of mind is concerned with the issue of free will. I do not believe that any device currently made by man has free will. I do not think that a thermostat can choose whether or not to turn on the central heating; I do not think that a driverless car has any choice in deciding where to drive to or how fast to drive; I do not believe that a fruit machine can deliberately decide not to pay out a winning jackpot. In fact, I think it is obvious that computational machines, however complex they may become in the future, can never exhibit what I understand to be free will.

If you accept my argument from evolution that free will exists in conscious brains then, obviously, brains are non-computational devices. The question now is, how can a brain *be* a non-computational device? How could such a device execute its own free will? What changes, if any, will we need to make in the laws of physics to allow a non-computational brain to execute its free will?

8 – Towards a non-Computational Theory of the Mind

The first thing to say is that, in order for an agent (e.g. a brain) to execute an action according to its free will, there must exist at least two possible potential actions which it can take and that all of these potential actions must be consistent with the laws of physics. Fortunately, many interpretations of Quantum Theory (QT) already allow for this possibility. Consider the following scenario commonplace in today's university laboratories:



A single photon of light from a photon source A is directed at a half-silvered mirror. QT predicts that there is an equal probability that it will be reflected or transmitted through the mirror. Some time later, when the photon has been detected at either the detector A or the detector B it becomes clear which of these alternative possibilities has become a reality but the curious thing is that, before the photon is detected it can be shown that the photon is in some sense in a superposition of states, having been both reflected *and* transmitted.

Now I do not wish to suggest that the photon has exercised its *free will* in choosing to be reflected rather than transmitted but you can see that there are remarkable parallels with what might be going on in a brain. Is it beyond the bounds of imagination to suppose that a brain, in choosing a course of action, is in some kind of superposition of states which ultimately collapses into a single state when the choice is realized?

Let us throw caution to the winds and make the following bold proposal:

When a conscious brain makes a decision between doing X or doing Y, parts of it at least are in a superposition of states and that QT could, in principle at least, determine the prior probabilities of the brain carrying out either X or Y. When the wave function which describes the state of the brain in this superposition collapses, the brain executes one of these options in strict accordance with those probabilities.

Because this statement is fundamental to my understanding of consciousness and free will I shall call it *the Quantum Hypothesis*.

Objections to the Quantum Hypothesis

One objection to this idea is that neuroscience has totally failed so far to fund any structures in the brain which look as if they could be in a superposition of states. That is not to say that QT is irrelevant to the operation of the brain. We already know that quantum effects are important in describing how neurons transmit electrical signals and quantum effects have been shown to play an important role in how plants photosynthesize and some animals detect magnetic fields. But nobody has (yet) identified those vital components of the brain which are uniquely responsible for our consciousness.

Roger Penrose and Stuart Hameroff have made a brave start in trying to identify structures within the brain that might be involved in this unknown quantum process which gives rise to consciousness and free will but but whether their ideas will bear fruit is still open to debate[39]. Their theory invokes the role of tiny structures inside neurons called microtubules whose function is obscure. One very interesting feature of the theory, however, is that anaesthetic drugs appear to have a profound effect on these structures without impairing the ability of the neuron to carry electrical signals. This strongly suggests that, whether or not the operation of these microtubules is quantum in nature, microtubules play a vital role in upholding consciousness and that all theories which ignore their role (which includes all the computational theories which regard neurons as classical logic gates) must be false.

The main objection to the idea that large parts of the conscious brain could be in a state of superposition, however, is this. QT as we currently understand it seems to apply only to objects on an atomic scale and only when great pains are taken to isolate the objects from the surroundings and keeping the temperature very low. The prospect of maintaining an object such as a brain weighing 2 kg at a temperature of 38°C in a superposition of states is thought to be utterly impossible. But the largest object which has currently (May 2023) been placed in a superposition of states is a sapphire crystal weighing 16 micrograms[40]. In addition, superconductivity – which is thought to be a quantum process – can occur at temperatures as high as 77 K (-196°C) so perhaps the idea that a brain, (or parts of it at least) can be in a superposition of states is not quite so unthinkable.

No one knows what causes the wave function to collapse (if, indeed, it does). Most interpretations of QT agree that it is something to do with the fact that when a system in a quantum

39 See, for example: *The Penrose-Hameroff Approach* https://www-physics.lbl.gov/~stapp/PenHam.pdf

40 https://www.scientificamerican.com/article/physicists-create-biggest-ever-schroedingers-cat/#:~:text=A%20sapphire%20crystal%20weighing%2016,superposition%20of%20two%20vibrational%20states.

superposition of states comes into contact with its environment it gradually looses its integrity – a process called decoherence – and it is argued that even if there are widely separated structures within the brain that are in a superposition of states, the coherence of the state would almost instantly be destroyed. But although decoherence may happen very quickly, it is not instantaneous. Now we often talk as if complex decisions take a long time to make; but it does not follow that the whole brain must be in a continuous state of superposition for a whole day when we are agonizing over whether or not to accept a new job offer. It may be that when a brain is conscious, messages are being sent all over the brain on what we might like to call the 'quantum telegraph' between 'quantum nodes' (microtubules?) which only remain in a state of superposition for microseconds.

There is some evidence for this sort of holistic activity in the brain. When we are asleep (or sleepwalking), different parts of the brain are doing many different things at the same time – monitoring our breathing, digestion, blood flow, sensory inputs from the nerves in the skin, ear and eyes etc. etc. But thinking a conscious thought seems to occupy the whole of our brain at the same time; you cannot consciously think two thoughts at once. While is is true that my grandmother was able to knit an Arran sweater while simultaneously watching Coronation Street and talking to her daughter, much of this behaviour was probably being handled at an almost unconscious level. If, during this activity, she heard the sound of breaking glass the whole of her conscious mind would switch to attending to a possible threat and the knitting would stop. It is possible that certain patients with brain disorders or so-called 'split brains' might be able to think conflicting thoughts but it seems likely that, whatever is the mechanism for conscious thought in the brain, it involves pretty well all of it at once.

This phenomenon has been noted by a number of proponents of the 'global workspace theory' mentioned on page 93 who point to the potentially massive interconnectedness of the brain to explain its apparent unity. I am not convinced by this. Each neuron in the brain is connected to as many as 7000 other neurons so, in principle, all 16 billion neurons in the human cortex could be connected together by a chain of as few as 4 neurons. However, this simple calculation is misleading because it is not the case that the cell can be connected to 7000 cells anywhere in the brain. A human neuron has a central cell called the soma which collects information from several thousand neurons in its immediate neighbourhood. This information is processed and integrated and sent along a single axon (which may indeed go to the other side of the brain) where it is connected to several thousand other neurons – again only in the immediate vicinity. So while it is true to say that the 3500 neurons at one end of the axon are 'connected' to the 3500 neurons at the other, they cannot send independent messages to one another. This bottleneck is bound to have serious consequences for the speed at which conscious computation can take place. Take reading, for example. It is normal to read a page of a novel containing 400 words in about a minute. The actual task of processing the visual information is done by specialist hardware in the eye and the visual cortex; but what takes time is the extraction of meaning from the symbols by the conscious brain. Suppose the visual cortex supplies the conscious brain with the word 'grandfather'. The reader must first recall what a grandfather is, then he must figure out whose grandfather the author is talking about and where this person fits into the story etc. etc. There is no specialised hardware in the brain for doing all of this; the whole conscious brain is involved in retrieving information concerning grandfathers in general and what has happened so far in the story from parts of the brain which must of necessity be several centimetres apart. If the brain was just a neural network with 16 billion nodes whose average connection time was a few milliseconds, then it would probably take about an hour to settle on the meaning of the word 'grandfather'!

But if there exists inside the conscious brain some sort of internal 'quantum telegraph' this problem is solved. For a split second, millions of 'quantum nodes' over large areas of the brain are connected together exchanging information far faster than the rate at which neurons can transmit electrical impulses.

Another objection to the idea that it is just massive interconnectedness which explains consciousness is that there doesn't seem to be any reason why this process should just 'switch off' when we are asleep or unconscious. It is also interesting to note that the power consumption of the brain only falls by about 15% during non-REM sleep. This would seem to indicate that, whatever process is responsible for consciousness, it does not require very much power. This rules out the possibility that during sleep it is just the long-range neurons which are switched off because if that were the case, we might expect the power consumption of the sleeping brain to be very much smaller than the power consumption of the conscious brain.

Let us continue our speculative journey by assuming that free will is a process by which a brain in a superposition of states realizes a single choice of action. What makes this process different from the process by which a photon chooses which way to go?

As you know, a brain is an assembly of billions of interconnected neurons. When a brain is coming to a decision these neurons can be divided into two classes. There are those neurons which store the information on which the decision must be based and there are those neurons whose activation is required for the brain to carry out its decision. Exactly the same is true of the circuits in the 'brain' of a Mars rover facing the choice of deciding whether to go to the left or to the right of a rock blocking its path. Some of the circuits contain information which encodes the instructions which the rover has received from Earth; other circuits tell it how big the rock is, how far away it is etc. There may even be circuits which contain information about what happened the last time the rover came across this particular rock. The program stored in the rover's memory executes a deterministic algorithm which sequentially sifts all this information and comes up with a number which integrates all this information and which, if positive, causes the rover to go left and if negative, go right. The rover may also possess a random number generator to decided which way to go if the number computes to zero.

We know for sure that human brains do not work in this way. At the very least they are massive neural networks which store their 'programs' in the way the neurons are connected rather than specialised 'memory hardware' But I have already pointed out that any neural network can be simulated by a conventional stored program (Turing) machine so this is not the reason why brains can execute free will but computers cannot. The difference is, I believe, that the conscious brain can integrate all the necessary information in the form of a 'wave function' which connects together all the relevant neurons simultaneously via what I have called the 'quantum telegraph'. When this wave function collapses, the outcome is that the relevant motor neurons are activated and the action is carried out.

Comparing this scenario with the photon and the half silvered mirror we could say that the process is fundamentally identical, the only difference being that the information the photon has access to is confined to a single 'random bit generator'. I suppose you could say, if you like, that the photon has one 'bit' of free will, but it has no 'bits' of consciousness.

Does the Mars rover have any 'bits' of consciousness? No. Because it is never in a superposition of states. Does the Mars rover have any 'bits' of free will? No. Because its response is either entirely determined by the states of its internal circuits or, in rare instances by its random number generator.

What about a brain? What is the crucial difference between the decision that a brain makes when deciding whether or not to accept a new job or the Mars rover in deciding whether to go left or right? The brain's response is definitely not random in the sense that when we make decisions we do take into account all sorts of factors. Suppose you have just received a letter offering you a new job. Before posting your reply accepting or rejecting the offer you will consider how much better the pay is, whether your wife will approve of the necessary move, what the schools are like in the area etc. etc. etc. If you really can't decide, you might toss a coin (human brains do not appear to possess a random bit generator. They are notoriously bad at inventing random numbers.)

The big question is whether, having integrated all the available information and come to a decision, the brain can truly be said to have made a free choice. The answer to this is a resounding YES. Unlike the Mars rover which integrated all the information available to it and chose to go left, that choice was entirely determined by the state of its memory circuits at the time and could have been predicted in advance by any computer engineer with access to its memory banks. The response of the human brain, however, could not be predicted in advance because it was in a superposition of states whose outcome can only be predicted by the laws of physics as a probability. In fact, if you were able to write down the equation describing the exact state of a human brain in the act of deciding whether or not to accept the offer of a job, QT will predict the probabilities of the two possible outcomes. If the pay is fantastic, the wife is enthusiastic and the schools brilliant, QT might even come up with the solution 99% accept the job, 1% reject it. In reality the choice is usually not so obvious and QT may only predict a much less helpful split split.

It is at this point that my materialist detractors jump in gleefully saying that all this means is that our difficult decisions are as random as the choice of the photon in deciding whether to reflect or transmit, or the choice of the uranium atom in deciding whether or not to decay. What they forget, however, is that this quantum process which weighs up the pros and cons of the different inputs is also responsible for the existence of a *conscious self*. From the point of view of QT the decision is indeed random in accordance with the laws of physics; but *from the point of view of the conscious self* the decision was made rationally.

Now we see clearly why consciousness is a necessary component of free will. The difference between the photon on the one hand and the brain on the other is simply that the latter is conscious and the former is not. It is not that the conscious brain has to doctor the quantum probabilities to get the result it wants, it is more a case of the conscious brain, having weighed up the options and made a decision consistent with the probabilities determined by QT, has to accept responsibility for that decision. In other words, conscious brains have a self and selves can be held responsible for their actions. Photons do not. Is this such a hard thing to accept?

Consider another example. I wake up in the night and go down to the kitchen to make myself a cup of tea, or perhaps a cup of coffee. As I approach the kitchen cupboard my conscious mind is processing all sorts of information which might include the price of coffee, how many tea bags there are left, how much coffee my wife will need for her coffee morning tomorrow, whether I actually like the brand of tea that I find in the cupboard etc. etc. As I lift up my hand to select the coffee jar or the tea caddy, many components of my brain are in a state of superposition from which there may spring many alternative histories and which encodes all that afore-mentioned information. When my conscious mind decides to reach for the tea caddy, the outcome is *not pre-determined* (because the laws of physics only predict the probabilities), *nor is it random* (because it comes about as a result of the state of superposition of certain elements in my conscious brain which encoded all the relevant information which are needed to make the decision). And when the tea is drunk, my conscious self has to take responsibility for the action and face the music when my wife finds that I have used up the last tea bag.

Contrast this with the behaviour of a sleepwalker who gets up in the middle of the night and makes a cup of tea. When he approaches the cupboard, his brain is not in a state of superposition. It is simply an automaton going through well-rehearsed motions. There is no thought in his mind as to whether there will be enough coffee left for his wife tomorrow or whether or not he actually likes the tea in the caddy; his action is wholly determined by habit and if the jar and the caddy had been placed the other way round, he probably would have made coffee instead, putting the coffee into the tea pot as well!

If this theory is correct then a non-computational or quantum brain is a necessary condition for free will and if we can prove that a creature is exercising its free will, then it must possess such a brain. If, on the other hand, you believe that the human brain is a computational device, then you must deny that humans have free will. This is, of course, a perfectly consistent philosophical position to adopt and has been the default assumption of classical determinism ever since the discovery of Newton's laws of motion. I, personally, cannot accept this. I simply have to believe that I have a choice in the matter of what I do next, otherwise there is absolutely no point in discussing moral issues, interviewing for a job, putting money into a bank, shopping for tomorrow's dinner, writing a book on consciousness or even taking another breath. What is more, a society in which interviewing for a job or writing books is a daily occurrence could not have come about by Darwinian evolution. Free will exists and if that fact is incompatible with the laws of physics as we currently understand them, then so much the worse for the laws of physics.

Quantum Theory and Sentience

The Quantum Hypothesis gives us some sort of hope that we might discover structures in the brain which enable us to exercise our free will but it does not help to explain how or why we are *sentient*. How any physical process in the brain could give rise to the subjective feeling which we all experience when we are conscious is what David Chalmers calls the 'hard problem[41] and I have no insights to offer on this issue. Indeed, I suspect that it will remain a mystery which will always remain beyond the wit of man to resolve. It may even turn out that there is a logical contradiction in the concept of a brain which fully understands itself just as there is a logical contradiction in the idea that a crane can lift itself up or the a box can contain within it an identical box.

From an evolutionary point of view, sentience seems to have developed first. Unconscious creatures can respond appropriately to the world around them but they do not perceive the world as being *out there* because they do not perceive themselves as being *within* it. According to my theory, the first conscious creatures used superposition and the 'quantum telegraph' to integrate incoming sensory information rapidly into a mental map of the world outside them. They developed conscious feelings and a sense of self. This led to the recognition that there were other creatures *out there* with a sense of self (*empathy*) resulting in the development of pair-bonding and of hierarchical societies.

The same process that produced sentience had other consequences. The first is long-term memory. Not only was it possible to place themselves is space, it was also possible to locate their selves in time as well. They found that they could remember what had happened in the past and predict what might happen in

---

41 David Chalmers: *Facing Up to the Problem of Consciousness* https://consc.net/papers/facing.pdf

the future. It became possible for them to have internal thoughts or what I have called *sapience*.

Finally, a few species discovered that they had a new power – the power of free will. Not only could they predict the future, they could change it. They discovered that they could imagine new ways of doing things and create things that had never been created before. The rest, they say, is history.

9 – Summary and Conclusions

So where does all this talk of superposition leave us? On the one hand I will freely admit that the idea is a mere suggestion and that it comes nowhere near to being a scientific theory. On the other hand, I am convinced that consciousness is not an emergent phenomenon – it is, in my view, a *transcendent* one and will need new physics to explain it.

In recent years there has been a plethora of books claiming to have 'explained consciousness' by appealing to '*Strange Loops*'[42], '*Multiple Drafts*'[43], '*Global Workspaces*'[44], '*The Ispundrum*'[45] etc. etc. – but all of these attempts assume that the (conscious) brain is just a massively parallel neural network. I am utterly convinced that this is not the case and if we are to make any progress at all in understanding how our conscious brains work, we must be prepared to contemplate some radical new ideas.

I have already listed on page 92 a number of questions which our current understanding of the laws of physics is unable to answer. It would, I think, be totally unreasonable to suppose that this list was anything like complete. The currently fashionable claim that we know everything there is to know about the workings of the brain at the neuron level but that we just don't know the details of how it is organised, is both arrogant and absurd. The mystery of consciousness and our sense of self and being is so profound and is so at variance with the laws of physics

---

42 Douglas Hofstadter: *I am a Strange Loop*

43 Daniel C. Dennett: *Consciousness Explained*

44 Stanislas Dehaene: *Consciousness and the Brain*

45 Nicholas Humphrey: *Sentience*

as we currently understand them that it is surely impossible to deny that we are missing something really important.

Of course it is at this point that my readers of a mystical persuasion (if there are any still reading that is!) will jump in triumphantly and say “At last! You have just twigged! What you are missing is – the immortal soul[46], élan vital[47], parallel temporal dimensions[48], fractal attractors[49]...” or whatever catchy phrase you want to use to explain the unexplained. I do not think we have to throw in the towel just yet. If there are processes going on in the brain which our current laws of physics seem to render impossible then it is our duty as scientists to find out what those processes are and reconcile our laws of physics to fit the facts, not deny the facts to fit our current theories..

It is worth comparing the attitude of modern neuroscientists in denying the possibility of a role of quantum theory in the workings of the conscious brain with that of cosmologists grappling with the with the problems of Dark Matter and Dark Energy.

Measurements of the Doppler shift in the frequency of light emitted by stars in distant galaxies tell us that the stars do not rotate in the way they ought to. The accepted inference is that galaxies contain a lot of matter that we cannot see. This has initiated a lot of effort into dreaming up the existence of all sorts of exotic forms of matter which no one has ever observed. In the case of Dark Energy, cosmologists are prepared to countenance fundamental changes in the laws of physics just to explain why a few different quasars are dimmer than they ought to be.

But when it comes to explaining consciousness and free will, the very suggestion that quantum processes could possibly be

46 Descartes: *Méditations Métaphysiques.* 1647
47 Henri Bergson: *Creative Evolution* 1907
48 J. W. Dunne: *An Experiment with Time* 1934
49 Tim Palmer: *The Primacy of Doubt* 2022

involved and that the brain could be a non-computational device is met with derision. Even if the Quantum Hypothesis is complete rubbish, just the mere recognition that the conscious brain is qualitatively different from the unconscious brain is a big step forward in understanding it and being able to recognise it in other creatures.

In our search to understand what consciousness is we have come across a number of different ways of recognising consciousness in other creatures. First there is *sentience*: any creature which is conscious has a sense of self and feels things including pain. In Nagel's terms, there is 'something it is like' to be a conscious creature.

The trouble with this is that sentience is, by definition, subjective and we currently have no way of positively identifying when a creature is sentient. It may be that, in the future, when we have discovered the secret of superposition – or whatever it is that distinguishes the sentient brain from the unconscious brain – we might be able to build a scanner which will detect the 'aura' of sentience, but at the moment we must fall back on behavioural cues of which the most important is what I call *empathy* – i.e. the ability of one conscious creature to recognise another creature as an individual and to relate to that individual in a unique way. It is this quality of empathy which enables many mammals to live in social groups and many birds to live with the same partner for a season or even for life.

Other indicators of potential sentience such as violent reactions to situations which in us would cause pain must be treated with extreme suspicion because all creatures have built-in mechanisms to avoid potentially harmful situations. On the other hand, if a creature displays empathy or any of the other indications of consciousness, then it will almost certainly feel pain.

The enjoyment of pleasure is a more reliable guide than pain because there does not seem to be an evolutionary benefit in doing things simply for the pleasure it brings. So if you see a puppy playing with a ball or a parrot tobogganing down a slope, you can be pretty sure they are conscious.

The second feature of consciousness which we can use to identify it in other creatures is *imagination*. Only non-computational brains can create something genuinely new so if you see an animal displaying some novel behaviour, it is almost certainly conscious. This is very rare, however, so it does not get us very far.

The third indicator is perhaps the most useful of all. I have argued that only non-computational brains have free will – and we can be pretty sure that consciousness is a necessary consequence of non-computability. It follows that creatures with free will must be conscious. Recognising and identifying free will in another creature is not straightforward, however. I have used the word *intention* to indicate behaviour which shows that the creature a) knows what it wants and b) knows what actions it must carry out to get what it wants based on memories of how it achieved similar goals in the past. The point about memory is important here. If you present a creature with a binary choice, one of which leads to a cache of food, the fact that it chooses route A and not route B is no indication of free will. You must demonstrate that the creature knows that route A will lead it to food – because it has often lead to food in the past, for example. A creature without memories of the past can never be said to exercise its free will because it has no grounds on which to base its decision.

Now I suggested on page 60 that it is only creatures which display other evidence for consciousness which appear to have long-term memories. This suggests to me that the process which is responsible for consciousness (e.g. superposition) might also have an important role to play in the way long-term memories are stored in the brain. Of course, I would not conclude that we therefore forget everything that we ever knew when we fall asleep; what I am suggesting is that long-term memories may be stored in a form which is only *accessible* to the brain when it is in a state of superposition. But if this is true, then if we see evidence for long-term memory (i.e. at least a few weeks) then we are probably looking at a conscious creature.

I must add one caveat here. Our current state of technology has enabled us to build machines such as the Mars rovers which in every way show evidence of *intention*. They act totally autonomously and appear to have complex goals and store information for long periods of time. So do cruise missiles. We can be sure, however that these devices are not conscious because we know exactly how they work. Evidence of intent is therefore no a reliable guide to the existence of sentience and should not be equated with free will.

Finally, in steering a delicate path between the Scylla of determinism and the Charybdis of randomness we lay ourselves open to arguments from above and below. The die-hard determinist argues that uncaused events such as the collapse of the wave function which describes my state of mind immediately before making a crucial decision are, almost by definition, random and therefore do not count as free will – conveniently forgetting that the probabilities that ultimately determine which course is chosen are determined by a *conscious* brain simultaneously taking into account all the factors that have a bearing on the decision. On the other hand, the dualist (or spiritualist) may argue that what I have proposed comes close to saying that there exists in the conscious brain something which exists outside the laws of physics and which can reach down and manipulate reality whenever it chooses to do so. In so far as I believe there there *is* something going on in the brain which we genuinely do not understand I might even be tempted to accept this point of view, but I would describe myself as a *mysterian* rather than a dualist. I see no reason at all to infer that consciousness is a phenomenon which is beyond the reach of scientific enquiry. And I certainly do not believe that there exists an entity in all of us which is capable of reaching down and manipulating the laws of physics whenever we make a decision. Consciousness is simply a mystery – and we are not going to settle the question of how we can have free will in spite of what scientists are currently telling us until we have a much better understanding of the nature of consciousness.

The answers to my seven questions

In conclusion therefore, what are my answers to the seven questions posed on page 13?

A) Are there degrees of consciousness?

Technically, the answer to this is no. Either the brain (or part of it) is in a state of superposition or it isn't. On the other hand, I have argued that consciousness only makes any sense in the light of past experiences, so newly born creatures or a human being with total amnesia cannot really be said to be conscious *of* anything. Also we could surely be forgiven for saying that a chimpanzee with 10 billion neurons in a state of superposition is in a sense more conscious than a mouse which only has 71 million cortical neurons.

B) What creatures other than human beings possess consciousness?

Probably all mammals and birds in some degree. Possibly cephalopods and maybe even a few fish. There is currently no evidence that any reptiles, amphibians, crustaceans, molluscs or any other living creatures including plants are in any way conscious.

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

Technically at about 32 weeks in the womb – but see my answer to A.

D) What are the evolutionary benefits of consciousness?

The most important one is the ability to recognise other creatures as individuals – what I have called *empathy*. The ability to 'think outside the box' (*imagination*) and the ability to exercise *free will* has had huge implications for the evolutionary development of humans but it has had little effect on the development of other species.

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

I believe it will be possible to identify the physical process or processes which give rise to consciousness but whether this will give us a satisfactory answer to Chalmers' 'hard question' is another matter.

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

Yes. In fact I would go so far as to say that, in my opinion, the 'age-old problem' is already solved. Free will exists and can plausibly be explained as a quantum phenomenon related to superposition.

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

In principle, yes. Once we have worked out how to put several billion quantum objects into a superposition of states for long enough for them to carry out a quantum calculation (or whatever it is that actually is responsible for consciousness) then we could perhaps build a conscious machine along those lines. But we should not underestimate the difficulties. The new physics involved will probably include solving the current incompatibility of Quantum Theory and General Relativity. Next, in order to make a machine which could do something useful (unlike a Bose-Einstein condensate which just sits there) we will probably have to figure out how to create or grow 16 billion quantum devices inside a cube of silicon or some other possibly biological material and then trigger it into action.

But then – my wife and I already know how to do that, so what would be the point?

C-FOS

The year is 1954

Alan Turing is sitting comfortably in the plush seat of a BOAC Stratocruiser on the long journey to Washington watching the clouds slide by 25,000 feet below. 6 weeks previously in March the world had been stunned – not to say thrown into a panic – at the arrival of a Unidentified Flying Object which had touched down in the Chihuahuan desert not far from the place where the first atom bomb had been detonated in Alamogordo, New Mexico. The landing site had been immediately sealed off by the US military and frantic conversations had ensued between President Eisenhower and Soviet Premier Khrushchev but nuclear war had, thankfully, been averted and, after an acrimonious debate in Congress as to whether or not the vehicle should be blown up immediately, the military were persuaded to settle for a period in which a close watch would be made for any activity within or outside the vehicle.

There were many eye-witnesses who had seen the craft as it descended, apparently without effort, on a column of blue light, but where the spaceship (for without doubt, that was what it was) had come from or how long it had been in the Solar System, nobody knew.

The world's press had confidently expected that any day a hatch would open and a troop of little green men would emerge chanting 'take me to your leader' but no such thing happened. For four weeks the craft just sat there in the sands, apparently doing nothing. Then a report came through from a radio ham in Los

Angeles claiming that he had detected a strange signal on a frequency of 106.37 MHz which he thought might be coming from the UFO and that, maybe, the UFO was transmitting signals back to a mother craft. This possibility was soon ruled out by virtue of the fact that the signals were, in fact, quite weak and would be unlikely to reach any great distance. It was also soon determined that the signal was frequency modulated but that only two frequencies were actually being used. In other words, the information that it contained was digital in character, not analogue.

Now the military radio engineers in charge of monitoring the signals were thoroughly familiar with frequency modulation and its use to transmit analogue signals but were stumped when it came to decoding this new type of signal. It was Edwin Bradbury, the director of the nearby Los Alamos National Laboratory who had suggested drafting in Jacob von Neumann from Los Alamos and Alan Turing from Manchester University, both experts in the emerging science of digital communication, to direct the effort to decode the signals emanating from the craft. This, then, was the reason Alan found himself 5 miles above the Atlantic ocean heading towards the strangest encounter of his, or anyone else's life.

When Alan finally reached the makeshift laboratory which had sprung up half a mile from the UFO he discovered that a small amount of progress had already been made. It appeared that the craft was repeatedly emitting a series of bursts of data never more than a millisecond in length. Using a 0 to stand for the lower of the two frequencies used and a 1 for the higher frequency, a typical burst could be written in binary like this:

...000001001010000000...

or

...0000011110000100000...

Bursts always started with a '1' which was 0.0873 ms long and the subsequent '1's were always the same length. The longest burst that the machine ever produced was 9 '1's long. It did not take Alan long to realise that the initial '1' was a marker and what followed was an 8 bit binary number.

Next, it appeared that these bursts were being transmitted in groups of 5 and, translating the binary into decimal notation, a typical group was:

37 64 42 73 79

which would be repeated 5 times before the group would change to something like

135 64 17 73 152

This was meat and drink to Alan who had spent 5 years of his life at Bletchley Park decoding far more complicated codes than this. He quickly realised that, in any group of 5 digits, the fifth number was always the sum of the first and the third. This meant, quite simply, that the second number (which was always 64) stood for + and the fourth (which was always 73) for =.

But why should a spacecraft be spouting simple addition sums to the world? Is it trying to say something? If that was the case, somebody suggested, why not answer back? Within the hour a small transmitter had been rigged up and, using the same carrier frequency, Alan's technicians began broadcasting some sums of their own. Instantly, the transmissions from the UFO ceased. Initially this was a great disappointment but it was soon discovered that the UFO had started transmitting on a slightly different frequency. There were still groups of 5 numbers repeated 5 times but this time the last number was the product of the first and third and the second number was always 66. When the penny dropped, it dropped with a clang that reverberated right around the world. 66 mean 'multiply' and the UFO was teaching us arithmetic!

When some journalist started to refer to the spacecraft as a Creature From Outer Space, it was inevitable that the object would henceforth be referred to as C-FOS.

Soon von Neumann and a team of mathematicians were quizzing C-FOS for all they were worth, trying to find out exactly how much mathematics it knew. After a while, it became clear that the machine was essentially familiar with the whole of number theory and was able to answer questions, expressed in the language of symbolic logic, such as 'is the number of primes infinite' and 'are there two integers *a* and *b* such that $a^2/b^2 = 2$' Neumann even asked a question that had remained unsolved for 200 years, namely 'are there any even integers which cannot be expressed as the sum of two primes'. The astonishing answer was ' yes; the smallest such number is …' and then C-FOS reeled out a number containing 8,379 binary digits. Von Neumann was staggered and immediately wired IBM with the request to check on the fastest computer they had available if this this astounding fact was really true. IBM complied but said that it would probably take several months, if not years, of computing time and would therefore cost thousands of dollars. Von Neumann replied that the job was too important and funding would be found whatever the cost. Meanwhile, headlines round the world screamed:

GOLDBACHS CONJECTURE FALSIFIED
MATHEMATICIANS STUNNED

Turing was not quite so impressed. 'OK' he thought to himself, 'so C-FOS can do impressive mathematics – but is it *thinking*? Is it *conscious*?' It was obvious that, in order to make real progress, an attempt must be made to communicate with the machine, not just through mathematical symbols, but through the medium of language. It was also obvious that, since the machine seemed to be, at least in some areas, vastly quicker and more intelligent (?) than a human, it would be easier for Alan to teach the machine English than to try to learn whatever language the machine might itself use to communicate with the other machines or aliens which, presumably, built it.

In order to make the task easier, Alan got his technicians to build what amounted to the first digital television. This consisted of a standard 525 line television tube of which the middle 461 were used for data. He then repeatedly broadcast a string of 212521 (= 461 × 461; 461 being a prime number) 8-bit numbers to C-FOS while simultaneously displaying the string as a greyscale image on the TV which was placed in front of what appeared to be a glass port on the side of the craft. The first image he used was a photograph of C-FOS itself. Almost immediately C-FOS responded with a string of 212521 numbers of its own. When this was fed into the TV it proved to be an image of Alan himself!

By showing C-FOS images of the solar system and the constellations, a limited vocabulary of astronomical terms was established from which it became clear the C-FOS had come from a planet in orbit round a star some 15 light years away. In response to images of planet Earth, its landscape, people and cities, C-FOS responded with images of its own planet. This appeared to be populated with a huge variety of machines obviously adapted to different purposes but all, apparently, built out of recognisably mechanical components. In none of the pictures was there anything which looked like some sort of master life form. It appeared that the whole planet was populated by – well, robots, and nothing else.

All this was, of course, quite fascinating; but Alan became increasingly frustrated by his inability to ask such simple questions as 'who built you?'. 'why did you come here?' and, of course, above all 'what do you intend to do now that you have got here?'

Over the succeeding years C-FOS proved to be unbeatable at chess but, though it always won, it often passed up obvious opportunities to force checkmate. It was as if it never really cottoned on to the *purpose* of the game. It proved several mathematical theorems including the famous Riemann Hypothesis but it never learned even the rudiments of English grammar. It could answer questions phrased as 'is X true?' but was totally unable to understand the question 'why is X true?'.

As interest gradually waned, a number of people started to suggest that C-FOS should be dismantled so that, at least, we could find out how it actually worked. Naturally Alan was vehemently opposed to this idea, though, by now, he had pretty well given up on the idea that the machine was conscious. It seemed to have no curiosity of its own, to have no imagination and very little purpose. Inevitably, though, his objections were overruled and it was decided to remove one of the panels which looked as if it could be an inspection cover. This was duly done and behind it there appeared to be a range of removable modules one of which proved to be a long metal box with what were obviously something like a million electrical contacts at the end. With the module removed, C-FOS's intellectual capabilities did not seem to be greatly impaired but the more modules which were removed, the more mistakes C-FOS seemed to make. Eventually when all 256 modules had been removed, C-FOS failed to respond coherently to any question it was asked but, as a machine, it was still clearly functioning and when the modules were restored, C-FOS was returned to his (its?) original state. Shortly afterwards, when all the technicians and scientists had gone home for the night, C-FOS took off. It was tracked for a while but it was too small to be visible for long and it was never established whether it (he?) was destined to go back to its home planet or whether it was off to find another planet to land on.

As to how it came to exist and why it landed on Earth, the most popular theory was that a race of conscious beings had, long ago, created a fleet of robots capable of building and maintaining themselves. For some reason, the conscious creatures had been wiped out, either by malicious intent on the part of the robots (this seemed unlikely considering the total absence of any suggestion of intent on the part of C-FOS) or, more likely, through some sort of pandemic. C-FOS was probably some sort of survey craft which had accidentally been thrown out of the stellar system in our direction.

The effect of the visit on humanity was both profound and minimal. On the one hand, it had proved that we were not alone in the universe in the sense that there existed intelligent creatures (or objects) out there – but it did not really answer the question 'are there any other *conscious* beings out there?' because the only creature we had encountered was clearly not conscious.

Mathematicians now knew that Goldbach's conjecture was false (which was interesting) and that the Riemann hypothesis was true (which was of vital importance) but since nobody could follow C-FOS's massive proof of the latter, the knowledge did not increase our understanding of the fundamental reasons which lay behind it one iota.

Ultimately, the first visit ever from a Creature From Outer Space was quietly forgotten and any suggestion that it had ever happened was met with a cascade of conspiracy theories and supernatural hocus-pocus. Only the mathematicians contemplating that 8,379 digit number which cannot be expressed as the sum of two primes remained convinced that C-FOS really had existed and had visited Earth in March 1954.

Bibliography

In researching the issue of consciousness and free will I have found the following books interesting if not always useful.

Kevin J. Mitchell *Free Agents* Princeton University Press (2023) 300 pages
Stance: Free will believer, all living things have some degree of free will
Style: Authoritative, wrong
Typical quote: “Living organisms [have a purpose]. That is, in fact, their defining characteristic.” (p166)
Catch phrase: Agency

Robert Sapolsky *Determined* The Bodley Head (2023) 404 pages
Stance: Free will denier
Style: Witty, informative
Typical quote: “It's turtles all the way down.” (p 82)
Catch phrase: Emergent complexity

Anil Seth *Being You* Faber and Faber (2022) 276 pages
Stance: Compatibilist
Style: Interesting but not entirely coherent
Typical quote: 'It is life, rather than information processing, that breathes fire into the equations.” (p 255)
Catch phrase: Conscious beast machine

Nicholas Humphrey *Sentience* Oxford University Press (2022) 217 pages
Stance: Computational mind theorist, free will is not discussed
Style: Excellent discussion of the evolutionary benefits of being conscious
Typical quote: “... circulating activity [in the brain] can be channelled and stabilized so that it settles into an 'attractor' state where a complex pattern repeats itself over and over.” (p 111)
Catch phrase: Ipsundrum

Peter Godfrey-Smith Other Minds William Collins (2016) 204 pages
Stance: Animals can be sentient without being conscious. (I disagree)
Style: Straightforward, non-technical, entertaining
Typical quote: “In the octopus's case there is a conductor, the central brain. But the players it conducts are jazz players, inclined to improvisation, who will accept only so much direction.” (p 105)
Catch phrase: <none>

Stanislas Dehaene *Consciousness and the Brain* Penguin Books (2014) 268 pages
Stance: Compatibilist
Style: Authoritative, closely argued
Typical quote: “Even if our brain architecture were fully deterministic, it would still be legitimate to say it exercises a form of free will.” (p 265)
Catch phrase: Global neuronal workspace theory

Mark Balaguer *Free Will* MIT Press (2014) 126 pages
Stance: The existence of free will is probable but unproven
Style: Straightforward, non-technical
Typical quote: “At least sometimes, when we come to make our torn decisions, nothing causes us to choose in the ways that we do.” (p 122)
Catch phrase: <none>

Douglas Hofstadter *I am a Strange Loop* Basic Books (2007) 363 pages
Stance: Consciousness is an emergent phenomenon
Style: Very entertaining with lots of ingenious ideas
Typical quote: “... our extensible repertoire of symbols gives our brains the power to represent phenomena of unlimited complexity and thus to twist back and engulf itself via a strange loop.”
Catch phrase: Strange loop

Donald R. Griffin *Animal Minds* University of Chicago (2001) 285 pages
Stance: This book is more of a comprehensive review of other peoples opinions rather than a statement of the author's views. (The bibliography alone runs to 43 pages)
Style: Comprehensive but inconclusive
Typical quote: I believe it is more likely than not that the emergent property of consciousness confers an enormous advantage by allowing animals to select the actions that are most likely to get what they want or ward off what they fear. (page 284)
Catch phrase: <none>

Colin McGinn *The Mysterious Flame* Basic Books (1999) 251 pages
Stance: Mysterian
Style: Rambling, inconclusive
Typical quote: It may be that the very faculties which make us so successful at the physical sciences unsuit our minds for understanding consciousness.
Catch phrase: Cognitive closure

Steven Pinker *How the Mind Works* Penguin Books (1998) 565 pages
Stance: Computational mind theorist, free will sceptic
Style: Penetrating but not always clear
Typical quote: "Our thoroughgoing perplexity about the enigmas of consciousness, self, will and knowledge may come from a mismatch between the very nature of these problems and the computational apparatus that natural selection has fitted us with." (p565)
Catch phrase: <none>

Daniel C. Dennett *Consciousness explained* Penguin Books (1991) 455 pages
Stance: Computational mind theorist, compatibilist
Style: Obscure
Typical quote: "All varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs." (p110)
Catch phrase: Multiple drafts

Index