Thoughts on Consciousness

David Mumford

Abstract: For many years, I have tried to come to a deeper understanding of what consciousness really means. This paper is a series of meditations about some aspects of this puzzle. First, I look at neuroscientists quest to *localize* consciousness in the brain and find wildly different conclusions. Second, believing that emotions are an essential component of consciousness, I explore some ideas of those who seek a theory of emotions. Third, I look at the connection of consciousness and time, especially the NOW. Fourthly, I try to lay out the pros and cons of which, if any, animals have consciousness. And finally, I discuss whether consciousness can find a home in silicon, in an "AI."

1. Introduction

Consciousness, as as essential quality of life, has always been a subject for philosophers and theologians, but beyond the purview of science. It is only recently that these taboos have been broken, that the word itself is not believed to be ineffable but that experts from a huge range of professions have felt it possible to say the word in print. Acknowledging that we are in deep water though, I like to quote Iris DeMent's song "Let the Mystery Be": "Everybody's wonderin' what and where they all came from," born with consciousness that then leaves us when "the whole thing's done." Well, I find "let the mystery be" hard to follow. I felt a bit enlightened when I read an op-ed piece in the New York Times [SK] that attempts to pin down the intuitive feeling of consciousness. Instead of the antiseptic word `consciousness', the author, Sean Kelley calls his piece "Waking up to the Gift of Aliveness." The article is a commentary on the sentence "The goal of life, for Pascal, is not happiness, peace, or fulfillment, but aliveness," that he traces in some form to his teacher Hubert Dreyfus. He confesses that he knows no explicit definition of aliveness but gives us two examples: looking at your lover's face when you have fallen in love; and lecturing to a class (he is a Professor) when your students are truly engaged and the classroom is buzzing. I take it that "aliveness" should be thought of as the most fully realized state of consciousness. While consciousness is the substrate of everything we do when we are alive in the mundane sense, the aliveness he is talking about is found in its most heightened moments, when all of life feels like it makes sense. He says aliveness should have the passion of Casanova without his inconstancy and the routine of Kant without his monotony. I'd like to think this is also the state of an enlightened Buddhist during meditation.

In this paper, I first want to review some neuroscientist's takes on consciousness and I want to look at recent work on emotions, arguing for the idea that feeling emotions is essential to consciousness. Then I want to look at consciousness and time, arguing that experiencing the flow of time is even more basic to consciousness. With all this background, I want to look closely at the question of whether and, if so, which animals possess consciousness in some form. Finally, I will review the present state of AI with a view to asking if a robot might ever possess consciousness.

2. Neuroscience and consciousness

The first time I encountered discussion of consciousness by any scientist was when I read the neuroscientist John Eccles' 1977 book [EP], joint with the philosopher Karl Popper, entitled *The Self and its Brain*. Both Popper and Eccles are believers in a *Three World* view of reality: (i) the objective physical world, (ii) the inner world of conscious beings and (iii) the world of ideas, that is objects of thought. Concerning the first two, they sought a detailed model of how in particular the physical brain interacts with conscious experience. Eccles developed their ideas further, e.g. in his 1990 paper [E2]. His hypothesis is first that the cerebral cortex can be broken up into about 40 million columnar clusters, each made up of about 100 pyramidal cells which stretch from near the inner to the outer cortical surface, clusters that he calls *dendrons*. Secondly, each dendron interfaces with a corresponding unit of conscious thought that he calls a *psychon* via an interaction allowed on the physical side by quantum uncertainty. A figure from [E2] are reproduced below. This is a breathtakingly bold and precise answer to the mind/body problem but one that has not drawn many adherents.

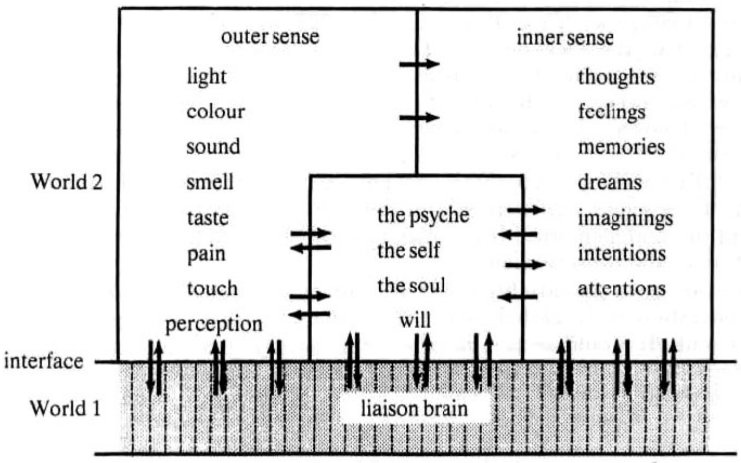

Figure 1. Eccles' diagram of the brain/mind interface from [E2]. World 1 is the physical realm, world 2 is consciousness,

More recently, scientists realized they could readily study *access consciousness*, that is whatever a person reports they are thinking, as opposed to consciousness in the ineffable, subjective sense of being alive, and then consciousness became something on which they could do experiments. Of course, they now exclude things like the reportedly heightened consciousness of Buddhists deep in meditation when all distracting thoughts of accessing the rest of the world are put aside. The goal of this research is to elucidate the *neural correlates of consciousness* with the aid of tools like fMRI, i.e. can one identify the precise neural states in which a person will report being conscious of something. It turns out that this is not as simple as one might hope: there are many sensations that cause measurable activity in primary sensory and other cortical areas that people are unaware of and there are neat experimental ways of producing them such as binocular rivalry, masking, and attentional distraction. Even the order in which two sensations occur can be experienced consciously as the opposite of their true order. Strikingly, the conscious decision to do an action seems to occur *after* there is brain activity initiating the action. Moreover, a certain amount of reasoning can be accomplished quite unconsciously by the brain. Freud would have told them that even strong emotions and the decision to perform actions resulting from these emotions often do not reach consciousness -- but his work was another taboo to scientists. The limitations of the self-awareness that consciousness provides were clearly summarized in Alex Rosenberg's NY Times piece op-ed piece *Why you don't know your own mind* [AR]. His conclusion is “Our access to our own thoughts is just as indirect and fallible as our access to the thoughts of other people. We have no privileged access to our own minds.”

What does make things conscious, according to many neuroscientists, is that the activity expressing some thought should spread over large parts of the brain, an idea known as the *global workplace theory* of consciousness, proposed first by Bernard Baars. This theory is almost the exact opposite of Eccles'. It proposes that activity over large parts of the cortex, often synchronized via ~40 hertz brainwaves called gamma oscillations, is necessary and sufficient for the thought to be conscious. Moreover each part of the cortex can contribute to the thought, including primary sensory areas but especially the pre-frontal cortex and the so called association areas of parietal cortex. I recommend Stanislav Dehaene's book *Consciousness and the Brain* for a detailed description of this theory. Dehaene writes on pp. 99-100:

> “Consciousness is like the spokesperson in a large institution. Vast organizations such as the FBI, with their thousands of employees, always possess considerably more knowledge than any single individual can ever grasp. ... As a large-scale institution with a staff of a hundred billion neurons, the brain must rely on a similar briefing mechanism. The function of consciousness may be to simplify perception by drafting a summary of

the current environment before voicing it out loud, in a coherent manner, to all other areas involved in memory, decision, and action."

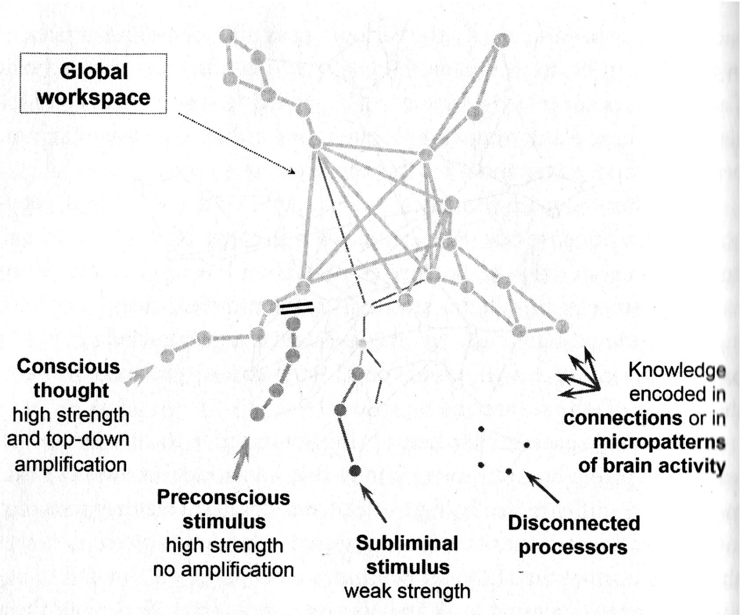
Global workspace
Conscious thought
high strength and top-down amplification
Preconscious stimulus
high strength no amplification
Subliminal stimulus
weak strength
Knowledge encoded in connections or in micropatterns of brain activity
Disconnected processors

Figure 2. Dehaene's diagram of conscious and unconscious activity in the brain and how a single thought ignites the workspce, from [SD], p.192.

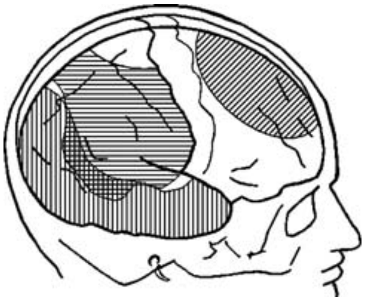

Figure 3. Large cortical excisions performed under local anaethesia by W. Penfield for the control of intractable epilepsy in three patients, entered on a single diagram. The patients remained conscious and communicative throughout the operation

But another theory for the location of consciousness goes back to Penfield's experiments operating on patients with epilepsy [WP]. To locate the exact cortical area whose excision would cure the epilepsy, he operated with local anesthesia and interrogated his awake patients while stimulating their exposed cortices on the operating table. In this way, he almost always found some area where the trigger for the epilepsy was located. But one form of epilepsy, *absence epilepsy* in which the patient briefly looses consciousness without any other symptoms, did not correspond to any unusual cortical electrical activity. This led him to propose that consciousness is *related not to the neocortex but instead to activity in the midbrain*. This theory has been extended by Bjorn Merker [BH] who filmed and worked notably with hydranencephalic children, children born with no neocortex (though the paleocortex and thalamus are usually preserved). His claim is that when given full loving custodial care and within the limits imposed by their many weaknesses, they exhibit behavior much like normal children.

For me, however, a really stunning piece of evidence for this theory was the `Turing test' of this issue that was carried out by Jaak Panksepp (see Panksepp's commentary to [BM], pp.102-103). He surgically removed the neocortex in 16 baby rats, paired them with normal rats and asked 16 of his students to each watch one such pair play and guess which rat was intact, and which lacked their neocortex. Only 25% of normals were correctly identified, while the decorticates were judged to be normals 75% of the time! It seems the major role of the neocortex was to make rats that possessed one more cautious, leaving the decorticates more playful. Panksepp's theory is that one specific midbrain area, the *periaqueductal gray* (PAG) (possibly together with its neighbors the ventral tegmental area (VTA) and the mesencephalic locomotor region (MLR)) gives rise to what he calls the *core self* or consciousness (see [PB], Chapter 11).

3. Is there a theory of emotions?

I believe that emotions are an essential ingredient of human consciousness. Unfortunately, the scientific study of the full range of human emotions seems stunted, largely neglected by many disciplines. For example, Frans de Waal, in his recent book *Mama's Last Hug* about animal emotions [FW], says, with regard to both human and animal emotions:

> "We name a couple of emotions, describe their expression and document the circumstances under which they arise but we lack a framework define them and explore what good they do."

One psychologist clearly pinpointed the role emotions play in human intelligence. Howard Gardner's classic book on his theory of multiple intelligences [HG] introduces, among a variety of skills, ``interpersonal intelligence" (chiefly understanding others' emotions) and ``intrapersonal intelligence" (understanding one's own). Psychologists have, of course, worked hard to define human intelligence. For a long time, the idea that they could pin this down with a numerical measure, the IQ, held sway. Does intelligence mean solving "Jeopardy" questions?; remembering more details about more events in your life?; composing or painting more skillful works? What Gardner realized is that intelligence includes many different skills but arguably what humans are uniquely good at, and what the large proportion of our everyday thoughts concern, is guessing what particular fellow human beings are feeling, what are their goals and emotions and even: what can I say to affect his/her feelings and goals so I can work with him/her and achieve my own goals? This is the skill that, more often than not, determines your success in life. This is now called "emotional intelligence" (EI) by psychologists but, as de Waal said, its study has been marred by the lack of precise definitions. A recent definition in Wikipedia's article on the EI is:

> "Emotional intelligence can be defined as the ability to monitor one's own and other people's emotions, to discriminate between different emotions and label them appropriately, and to use emotional information ... to enhance thought and understanding of interpersonal dynamics."

The oldest theory of emotional states is due to Hippocrates: *the four humors*, bodily fluids that correlated to four distinct personality types and their characteristic emotions. These were: sanguine (active, social, easy-going), choleric (strong willed, dominant, prone to anger), phlegmatic (passive, avoiding conflict, calm) and melancholic (brooding, thoughtful, can be anxious). They are separated along two axes. The first axis is extravert vs. introvert, classically called warm vs. cold with sanguine/choleric being extraverted, phlegmatic/melancholic being introverted. The second axis is relaxed vs. striving, classically called wet vs. dry, sanguine/phlegmatic being relaxed, choleric/melancholic always seeking more in life. This classification was developed in recent times by Hans Eysenck, whose version is shown here.

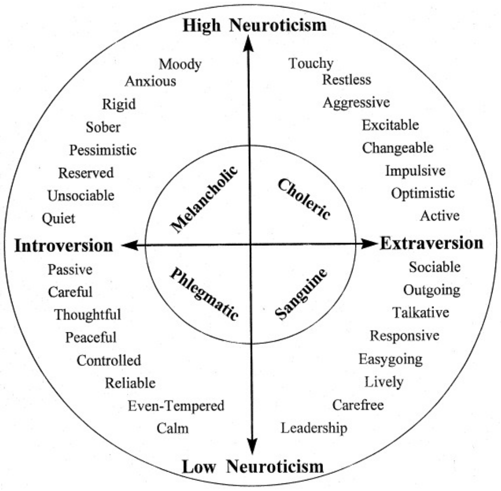

Figure 4. Hans Eysenck's axes and qualities elaborating the four humors.

The modern study of emotions goes back to Darwin's classic book *The Expression of the Emotions in Man and Animals* [CD] where he used the facial expressions that accompany emotions in order to make his classification. His theories were extended and made more precise by Paul Ekman and led to the theory that there are six primary emotions each with its distinctive facial expression, Anger, Fear, Happiness, Sadness, Surprise and Disgust and many secondary emotions that are combinations of primary ones, with different degrees of strength. Robert Plutchik has extended the list to eight primary emotions, named weaker and stronger variants and some combinations, resulting in this startling and colorful diagram:

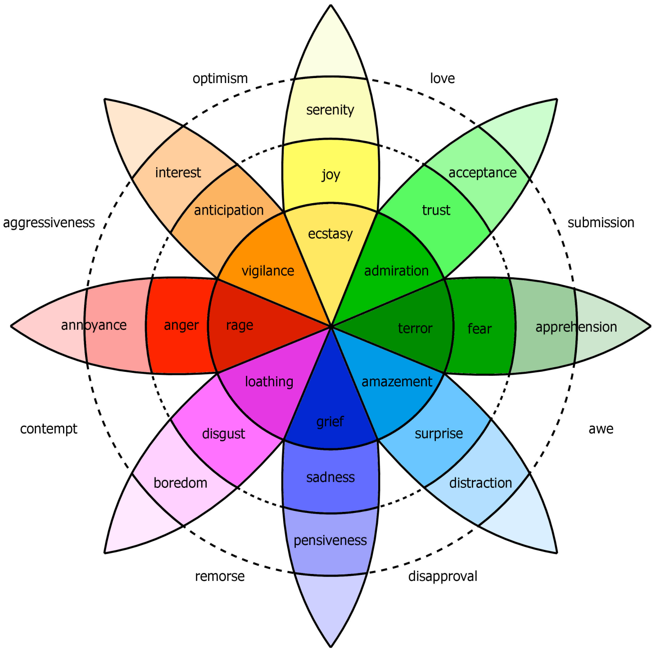

Figure 5. Robert Plutchik's elaboration of Darwin and Ekman's classification of emotions.

There is an open ended list of secondary emotions, e.g. shame, guilt, gratitude, forgiveness, revenge, pride, envy, trust, hope, regret, loneliness, frustration, excitement, embarrassment, disappointment, etc., etc. None of them to be just blends but rather grafts of emotions onto social situations with multiple agents and factors intertwined. Frans de Waal ([FW], p.85), referring to the above list, defines emotions by:

> "An emotion is a temporary state brought about by external stimuli relevant to the organism, It is marked by specific changes in body and mind -- brain, hormones, muscles, viscera, heart, alertness etc. Which emotion is being triggered can be inferred by the situation in which the organism finds itself as well as from its behavioral changes and expressions."

A quite different approach has been developed by Jaak Panksepp in [PB]. Instead of starting from facial expressions, his approach is closer to the Greek humors. Panksepp for a long time has been seeking patterns of brain activity (especially sub-cortical activity and the different neuro-transmitters sent to higher areas) that lead to distinct *ongoing affective states* and their corresponding activity patterns. This is different from emotions and the list is quite different from Darwin's though partially overlapping. He identifies 7 primary affective states: 1. seeking/exploring, 2. angry, 3. fearful/anxious, 4. caring/loving, 5. sad/distressed, 6. playing/joyful, 7. lustful.

An aside: I am not clear why he does not add an 8th affective state: painful. Although not usually termed an emotion, it is certainly an affective state of mind with sub-cortical roots, a uniquely unwelcome feeling and something triggering specific behaviors as well as causing specific facial expressions and bodily reactions.

No wonder de Waal said that as yet there is no definitive framework for emotional states. Perhaps what is needed to make a proper theory, usable in artificial intelligence code as well as science, is to start with massive data, the key that with neural networks now unlocks so much structure in speech and vision. The aim is to define three way correlations of (i) brain activity (especially the amygdala and other subcortical areas but also the insula and cingulate area of cortex), (ii) bodily response including hormones, heart beat (emphasized by William James as the core signature of emotions) and facial expression and (iii) social context including immediate past and future activity. An emotional state should be defined by a cluster of such triples -- a stereotyped neural and bodily response in a stereotypical social situation. To start we might collect a massive dataset from volunteers hooked up to IVs and MRIs, listening to novels through headphones. I am reminded of a psychology colleague whose grad students had to spend countless hours in the tube in the wee hours of the night when MRI time was available. Like all clustering algorithms, this need not lead to one definitive set of distinct emotions but more likely a flexible classification with many variants.

4. Time and the moving present

From my point of view, however, I think all scientists are missing the essential nature of consciousness. Sure we are conscious of what our eyes see and our ears hear, sure we are conscious of moving our body and making plans actions and sure, we can even fill our consciousness with the imaginary world of a novel or the proof of a theorem; but I think all this misses what makes consciousness absolutely different from anything material. Consciousness creates for us the experience of time, living in the present moment, the *now,* and it does this continuously moment after moment. I can hear a common response: isn't this just a part of the physical nature of the world?

To answer this, I want to quote from the two most famous physicists in order to clarify this idea. First of all, Newton, in his masterwork, *Principia*, states (translation from [IBC]):

> "Absolute, true, and mathematical time, of itself, and from its own nature flows equably without regard to anything external."

OK, this is indeed a good description of what time with its present moment feels like to us mortals. We are floating down a river -- with no oars -- and the water bears us along in a way that cannot be changed or modified. But now Einstein totally changed this world view by introducing a unified space-time whose points are events, each with a specific location and specific time. He asserted that there is no physically natural way of separating the two, no way to say two events are simultaneous when they occur in different places or that two events took place in the same location but at different times. Therefore, there is nothing in physics that corresponds to Newton's time. Not just that but science, essentially by definition, only studies correlations between events that can be reproduced exactly enough to show something is repeating itself, that here is a law of nature valid at least throughout some region of space-time. Thus it refuses to deal with any special instant that someone might call "now." The word "now" only enters our vocabulary through our conscious experience. For us, an experience is *never* reproducible (though we often try to make it so). As the saying has it, "you only go round once." Sure, physics studies unique events such as the explosion of the Crab nebula seen on Earth in 1054 CE, but this is a fact of history, not a scientific law. The science of astrophysics explains this explosion by equations which are then applicable to infinitely many stars and in this way removes the historical uniqueness of that supernova.

I want to mention an alternate theory, proposed by Richard Muller [RM] which recreates Newton's view. His 4-dimensional space-time has a boundary, called the *now,* and is growing as an objective universal time moves on, its past fixed but its future being continuously created. Such a theory, however, gained no traction with Einstein.

Einstein was fully aware that people experience what Newton was describing and wondered if this, with its notion of the present, could somehow have a place in physics. Though he never wrote about this, he had a conversation with Rudolf Carnap in the early 50's in which he made this point. (My thanks to Steven Weinstein for telling me about this conversation.) Here is how Carnap ([RC]) described it:

> "Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future, but that this important difference does not and cannot occur within physics. That this experience cannot be grasped by science seemed to him a matter of painful but inevitable resignation. He suspected that there is something essential about the Now which is just outside of the realm of science."

It's interesting to recall a famous debate between Einstein and the philosopher Henri Bergson (thanks to P. Mumford for this information). They met in 1922 arguing about the nature of time. For Bergson, the important notion of time was not that of clocks but that of people's immediate conscious experience (*Les Donnés Immédiates de la Conscience* was the title of his dissertation). When you focus on the subjective time of a person, it is indeed not bound up with his spatial location. I find his ideas hard to follow, but I think the key one is that time is heterogeneous, not homogeneous. Each instant for a conscious being is a thing in itself and their totality cannot be counted like a flock of sheep. Time, he says is a temporal heterogeneity, in which "several conscious states are organized into a whole, permeate one another, [and] gradually gain a richer content" (Stanford Encyclopedia of Philosophy). In contrast, at that point in his life, I assume that Einstein would have said that a person's lifetime is a curve in space-time, time-like meaning the person moves more slowly than light, bounded by the space-time points representing his birth and his death, and along which integrating the Lorentz metric computes each person's subjective time. You can see these guys are not going to reach a consensus. Apparently Bergson's denial of Einstein's theory of time was the reason his Nobel Prize was awarded instead for his work on the photo-electric effect so the meeting must have been tense. But the above quote concerning Einstein seems to suggest that later in his life he acknowledged some part of Bergson's point of view.

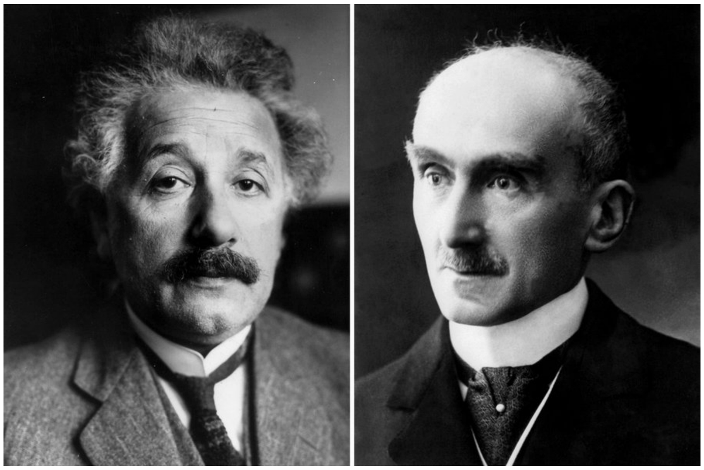

Figure 6. Einstein and Bergson about the time of their debate on the nature of time.

I find philosophical writings like Bergson's quite difficult to follow. But maybe one thing he is saying is that this experience of a present instant, the *now* that is always changing yet is always our one and only unique present, the one that each of us owns, that this *now* is the real core of what we call consciousness. I don't believe that sentience, the act of sensing the world, responding to these sensations, and the corresponding brain activity, are the most essential feature of consciousness. I believe that an experienced Buddhist meditator can put his or her self in a state where they wipe their mind clean of thoughts and then experience pure consciousness by itself, free of the chatter and clutter that fills our minds at all other awake times. Accepting this, consciousness must be something subtler than the set of particular thoughts that we can verbalize, the bread and butter of lab experiments on consciousness (e.g. Dehaene's work). In my few attempts at meditation, I found some measure of mental peace and quiet that gave me some insight into its potential. The idea that experiencing the flow of time is the true core of consciousness is the central theme of Eckhart Tolle's *The Power of Now* [ET]. The continual, ever changing, fleeting present is something we experience but that no physics or biology explains. It is an experience that is fundamentally different from and more basic than sentience and is what makes us conscious beings.

There is an important interaction between the conscious experience of time and the way emotions permeate consciousness. The fear of death and an intense will-to-live are among the most universal human experiences and clearly integrate the two aspects of consciousness that we explored in the last two sections. And, of course, we share these feelings with very many animals as I will propose in the next section.

5. Animal consciousness

Let me start by saying to whomever may be reading this paper: *I believe that you, my friend, do have consciousness!* Except for screwy solipsists, we all accept that "inside" every fellow human's head, consciousness resides that is not unlike one's own consciousness. But in truth, this is really a leap of faith because *we have no hard evidence for this besides our empathy*. So should we use empathy and extend the belief of consciousness to animals? Arguably, people with pets like dogs and cats will definitely insist that their pet has consciousness. Why? For one thing, they see behavior that is immediately understood as resulting from similar emotions to ones that they themselves have. They find it ridiculous when ethologists would rather say an animal is displaying "predator avoidance" than say it "feels fear." They don't find it anthropomorphic to say their pet "feels fear," they find it common sense and believe that their pet not only has feelings, but also consciousness. Our language in talking about these issues is not very helpful. Consider the string of words: emotion, feeling, awareness, consciousness. Note the phrases: we "feel emotions," we are "aware of our feelings," we say we possess "conscious awareness," phrases that link each consecutive pair of words in this string. In other words, standard English phrases link all these concepts and make sloppy thinking all too easy. One also needs to cautious: in our digital age, many elderly people are being given quite primitive robots or screen avatars as companions and such patients find it easy to mistakenly ascribe true feelings to these digital artifacts. So it's tempting to say we simply don't know whether non-human animals feel anything or whether they are conscious. Or we might hedge our bets and admit that they have feelings but draw the line at their having consciousness. But either way, this is a stance that one neuroscientist, Jaak Panksepp, derides as *terminal agnosticism*, closing off discussion on a question that ought to have an answer.

It is only recently that emotions as well as consciousness have gained the status of being legitimate things for scientific study. In the last few decades animal emotions have been studied in amazing detail through endless hours of patient observation as well as testing. Both Frans de Waal's book [FW] and Jaak Panksepp's book [JP] detail an incredible variety of emotional behavior, in species ranging from chimpanzees to rats and including not just primary emotions but some of the above secondary emotions (for instance, shame and pride in chimps and dogs). Panksepp has shown that young rats are ticklish and show analogous reactions to those of human babies when their bellies are tickled (see [JP] p.367). For me, these books and many others and, of course, my own meagre experiences with owning dogs, pigs, and watching horses and zoo animals makes a totally convincing case for animal emotions. Given the extensive organ-by-organ homology of all mammalian brains, I see no reason to doubt that all mammals experience the same basic emotions that we do, although perhaps not so great a range of secondary emotions. And if we all share emotions, then there is just as much reason to ascribe consciousness to them as there is to ascribe consciousness to our fellow humans. This is a perfect instance of "Occam's Razor": it is by far the simplest way to explain the data.

Going beyond mammals, it is useful to review the various stages of life, both living today and reconstructed from fossils, with a view to their potential for consciousness. I am inspired in doing this by the book *Other Minds: the Octopus, the Sea and the Deep Origins of Consciousness* [PGS] by the philosopher and diver, Peter Godfrey-Smith. At the base of the tree of life, we have two superficially similar kingdoms, the Bacteria and the Archaea. Both are prokaryotes, that is, are simple cells without nuclei, mitochondria, ribosomes or other organelles. On the other hand, both already possess proteins from the majority of protein families, as well as the universal genetic code (implemented by the same set of tRNA molecules) and, very significantly, they use the same complex electro-chemical mechanism as all higher life to synthesize ATP, their energy source. This mechanism uses ion pumps that make the cell membrane into a capacitor, the same mechanism that is used in higher animals as the key to information transmission in nervous systems (vividly described in Nick Lane's book [NL]). These simplest forms of life also sense their environment chemically via channels in their membranes and most can move in various directions using their flagella, thus reacting and seeking better environments. This is the beginning, a primitive form of *sentience* that started up c. 3.5 bya (billion years ago). It is perfectly possible that a mite of consciousness resides in these cells, but on the negative side, it is hard to see any sort of emotion in prokaryote existence.

The next step was the formation of much much bigger, more complex single celled organisms, the *eukaryotes* c. 2 bya. It is hypothesized that they started from an archaeon swallowing a bacterium, the bacterium becoming the mitochondrion in this new organism and, by folding its membrane again and again, hugely expanded the cell's ATP factory, hence its source of energy. Its skills sensing and moving got significantly better and, in order to control its shape and motion it developed *microtubules*. Roger Penrose and Stuart Hameroff have developed a controversial theory that quantum effects in these tubules in nerve cells create a vastly more powerful substrate for thinking and consciousness [HP]. I'm not aware of any other change that might have brought it closer to consciousness. But after that, around 0.65 bya (or 650 mya), multi-cellular animals formed. These were larger and obviously needed significantly better coordination, better senses and better locomotion. It is believed that the first nervous systems arose almost immediately to coordinate the now complex organisms. These creatures were soft and left almost no fossils but modern day jellyfish and sponges may be similar to organisms of that time. Sponges do not have nervous systems but jellyfish (and comb jellies) do and are the simplest organisms with nervous systems today. The environment is described as a mat of microbial muck covering the bottom of a shallow sea over which jellyfish like creatures grazed. Anyone for consciousness in this world?

The world becomes much more recognizable with the advent of predation, bigger animals eating smaller ones and almost all growing shells for protection, all this in the Cambrian age 540-485 mya. Note that predation introduces the experience of pain and the associated reaction, the will-to-live, and this must have been a nearly universal aspect of life in the Cambrian. Now we find the earliest vertebrates with a spinal cord. But we also find the first arthropods with external skeletons and the first cephalopods, predators in the phylum mollusca who grew a ring of tentacles and who, at that time, had long conical shells (see below an image of a reconstruction of the cephalopod *Orthoceras* from the following Ordovician age). In all three groups, there are serious arguments for consciousness. One approach is based on asking what animals *feel pain* and believing that feeling pain implies consciousness. There are experiments in which injured fish have been shown to be drawn to locations where there is a pain killer in the water, even if this location was previously avoided for other reasons. And one can test when animals seek to protect or groom injured parts of their bodies: some crabs indeed do this whereas insects don't. ([PGS], pp. 93-95 and references in his notes). Unfortunately, this raises issues with boiling lobsters alive, an activity common to all New Englanders like myself. In any case, inferring consciousness from the apparent perception of pain is an important idea. Another approach to inferring consciousness is the mirror test -- does the animal touch its own body in a place where its mirror image shows something unusual. Amazingly, some ants have been reported to pass the mirror test, scratching themselves to remove a blue dot that they saw on their bodies in a mirror (see image below).

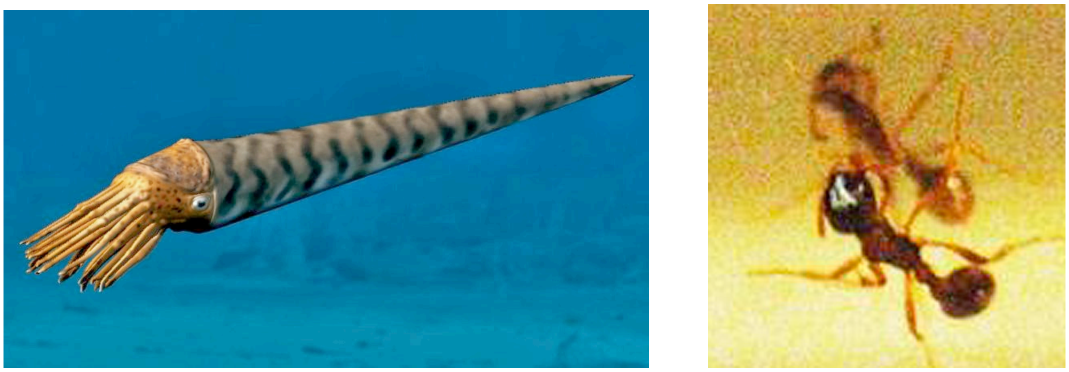

Figure 7. Left: the Ordovician cephalopod Orthoceras; Right: an ant trying to remove a blue dot that it sees in the mirror, from [CC].

.

With octopuses, we find animals with brain size and behavior similar to that of dogs. Godfrey-Smith quotes the second century Roman naturalist Claudius Aelianus as saying “Mischief and craft are plainly seen to be characteristic of the octopus.” Indeed, they are highly intelligent and enjoy interacting and playing games with people and toys. They know and recognize individual humans by their actions, even in identical wetsuits. As well as Godfrey-Smith's book, one should read Sy Montgomery's best seller *The Soul of an*

Octopus: A Surprising Exploration into the Wonder of Consciousness. Their brains have roughly the same number of neurons as a dog, though, instead of a cerebellum to coordinate complex actions, they have large parts of their brains in each tentacle. This is not unlike the way humans use their cerebral cortex in a supervisory role in any practiced action, letting the cerebellum and basal ganglia take over the detailed movements and simplest reactions. If you can read both these octopus-related books and not conclude that an octopus has just as much internal life, as much awareness and consciousness as a dog, I'd be surprised. The most important point here is that there is nothing special about vertebrate anatomy, that consciousness seems to arise in totally distinct phyla with no common ancestor after the Cambrian age.

Figure 8. Smart non-mammalian animals. Left: octopus playing with Mr. Potato Head; Right: crow using a 'tool'.

As mentioned, all mammals have virtually identical brains, differing only in the size of its constituent parts. Thus human brains are primarily distinguished by having a greatly enlarged pre-frontal cortex that appears to parallel the greatly increased planning activity carried out in our minds. On the other hand, non-mammalian vertebrates have brains which are all fairly similar to each other and similar to the mammalian brain *if you remove the neocortex*. The neocortex has a unique 6-layered structure not found in non-mammals although most recent thinking is that its parts are present, just not assembled and stitched together by pyramidal cells (as in Eccles' theory) as they are in mammals. These parts, called the pallium in birds and just the cerebrum in all classes, the 3-layered paleocortex, especially the hippocampus, as well as the thalamus (sometimes considered as a seventh layer of neocortex) are found in other vertebrates. Most importantly, the midbrain is present in all vertebrates and occupies a much larger percentage of the brain as you go down the evolutionary tree.

So how far down this tree can one make a convincing case for consciousness? Access consciousness cannot be used to interrogate animals without speech and, in any case, it hardly captures our experience of aliveness. With monkeys, the main experimental tool has been to train them to respond to a stimulus in different ways, e.g. by pressing various buttons, assuming that producing such a response means that the stimulus has activated something we can call their consciousness. In this way, a whole body of research has confirmed that consciousness in monkeys follows patterns similar to that in humans. For example, some stimuli do not reach consciousness and when they do, large parts of the monkey's neocortex show activity, often synchronized by gamma waves. But if Merker and Panksepp are right in believing that consciousness arises from midbrain structures, then all vertebrates should be considered conscious. Many people are convinced that birds, especially parrots and crows, are conscious beings and they lack a neocortex. And we have mentioned that reacting to injury or mirror images of bodily changes argues for consciousness in arthropods.

My personal view is that all the above also suggests that consciousness is *not* a simple binary affair where you have it or you don't have it. Rather, it is a matter of degree. This jibes with human experience of levels of sleep and of the effects of many drugs on our subjective state. For example, the anesthetic versed (midazolam) creates a bizarre half conscious/half unconscious state of which you have no memory. As our brains get bigger, we certainly acquire more capacity for memories but some degree of memory has been found for example in fruit flies. When the frontal lobe expands, we begin making more and more plans, anticipating and trying to control the future. But even an earthworm anticipates the future a tiny bit: it "knows" that when it pushes ahead, it will feel the pressure of the earth on its head more strongly and that this not because the earth or some other animal is pushing it backwards, i.e. they anticipated the push back ([PGS], p.83). My personal belief again is that some degree of consciousness is present in all animals with a nervous system. On the other hand, Tolkien and his Ents notwithstanding, I find it hard to imagine consciousness in a tree. I have read that their roots grow close enough to recognize the biochemical state in their neighbors (e.g. whether the neighbor tree is being attacked by some disease) but it feels overly romantic to call this a conversation between conscious trees.

6. Can computers be conscious?

The theory of artificial intelligence, during my lifetime, has gone through half a dozen cycles of boom and bust: periods when it was said confidently that computers will soon attain human level intelligence and periods of disillusionment when this seemed nearly impossible. Today, we are in the latest boom period and some visionary computer scientists are going even further, asking when "AI"s (using the abbreviation AI to make the computer sound like a new life-form) will actually attain not merely human intelligence but possess our consciousness as well. Other futurists ask for a wilder, crazier life altering boon: can I live forever by having my brain and my consciousness downloaded into silicon, essentially a person metamorphosing into an AI? In the boom of a previous cycle, the wild prediction was that we were headed for "the singularity," a point in time when super-AIs will create a wholly new world that leads to the extinction of the now superseded human race (predicted by some to happen around 2050). I plead guilty to personally hoping myself, half a lifetime ago, that I would be a witness the first computer attaining consciousness. But now I am quite a bit more skeptical. Perhaps this is the negativity of old age but perhaps too it is because I see this question as entraining issues not only from computer science but also from biology, physics, philosophy and, yes, from religion as well. Who has the expertise to work out how all this impacts our understanding of consciousness? I'm aware that even talking about the relevance of religion to any scientific advance is anathema to today's intelligentsia. But just consider this: is there a belief system in which the Silicon Valley dream that humans will soon be able to live forever and the Christian credo of "the immortality of the soul" are both true? For me, these two beliefs live in separate universes.

Let me add some further comments on the present AI boom and why it may lead to a bust in spite of its successes. The central player in the codes that support the new AI is based on an algorithm called a *neural net*. Every net, however, uses zillions of parameters called its *weights* that must be set before it can do anything. To set them, it is ``trained" using real world datasets by a second algorithm called *back propagation*. The resulting neural net then takes a set of numbers representing something observed as its input and it outputs a label for this data. For example, it might take as input an image of someone's face represented by its pixel values and output its guess whether the face was male or female. Training such a net requires feeding the net with a very large number of both male and female faces correctly labeled male or female and successively modifying the weights to push it towards making better predictions. Neural nets are a simple design inspired by a cartoon version of actual cortical circuits that goes back to a classic 1943 article of McCulloch and Pitts [MP]. More importantly, in 1974, Paul Werbos wrote a PhD thesis [PW] introducing back propagation in order to optimize the huge set of weights by making them work better on a set of inputs, i.e. a dataset that has been previously labeled by a human. This was played with for 40 years and promoted especially by Yan LeCun with some success. But statisticians were skeptical it could ever solve hard problems because of what they called the bias-variance trade-off. They said you must compare the size of the dataset on which the algorithm is trained to the number of weights that must be learned: without enough weights, you can never model a complex dataset accurately and if there are enough weights, you will model peculiar idiosyncrasies of your dataset that aren't going to be representative of new data. So what happened? Computers got really fast so neural nets with vast numbers of weights could be trained and datasets got really large thanks to the internet. *Mirabile dictu*, in spite of statistician's predictions, the algorithm worked really well and somehow, magically avoided the bias-variance problem. I think it's fair to say no one knows how or why they avoid it. This is a challenge for theoretical statisticians. But neural nets are making all kinds of applications really work, e.g. in vision, speech, language, medical diagnosis, game playing, applications previously thought to be very hard to model. To top it off PR-wise, training these neural nets was now been renamed *deep learning*. Who could doubt that the brave new world of AI has arrived?

BUT there is another hill to climb. In a blog post entitled "Grammar isn't merely part of language" [DM], I discussed the belief that thinking of all kinds requires grammars. What this means is that your mind discovers patterns in the world that repeat though not necessarily exactly. These patterns can be visual arrangements in the appearance of objects, like points in a line or the position of eyes in a face, or they can be the words in speech or simple actions like pressing the accelerator when driving or even abstract ideas like loyalty. No matter what type of observation or thought carries the pattern, you expect it to keep re-occurring so it can be used to understand new situations. As adults, everything in our thoughts is built from a hierarchy of the re-usable patterns we have learned and a full scene or event or plan or thought can be represented by a "parse tree" made up from these patterns. But here's the rub: in its basic form, a neural net does not find new patterns. It works like a black box and doesn't do anything except label its input, e.g. telling you "this image looks like it contains a face here." In finding a face, it doesn't say – "first I looked for eyes and then I knew where the rest of the face ought to be." It just says tells you its conclusion. We need algorithms that output: "I am finding a new pattern in most of my data, let's give it a name." Then it would be able to output not just a label but a parse of the parts that make up its input data as well. Related to this desideratum, we are able to close our eyes and imagine what a car looks like, with its wheels, doors, hood etc. That means we can synthesize new data without going out in the world to find it. This is like running the neural net backwards, producing new inputs for each output label. Attempts to soup up neural nets to do this are ongoing but are not yet ready for prime time. How hard it is to climb this hill is an open question but I think we cannot get near to duplicating human intelligence until this is solved.

I have argued above that what humans are uniquely good at is guessing what particular fellow human beings are feeling, what are their goals and emotions. Computer scientists have indeed looked at the non-emotional, logical side of modeling other "agent's" knowledge and plans. A helpful example that illustrates some of the subtlety of two agents cooperating is given by imagining two generals A and B on opposite mountain tops needing to attack an enemy in the valley between them simultaneously but only able to communicate by sneaking across enemy lines. A sends a message to B: "attack tomorrow?," B replies "yes." But B doesn't know his reply got through and so A must send another message to B that he did get the earlier message and knows B will act. But if this message hadn't gotten through, B would still be in doubt about A's knowledge, so B has to send yet another message to A. In fact, there is no end to the messages they need to send to achieve full common understanding. Computer scientists are well aware that we need to endow their AI with the ability to maintain and grow models of what all the other agents in its world know, what are their goals and plans and even what they know about what you know. This must include knowing what they themselves don't know. But arguably, this is all do-able with contemporary code.

However, as I said above, an essential ingredient of human thought is missing: emotions and emotional intelligence. To my knowledge, the only computer scientist so far who has endeavored to model emotions, is Rosalind Picard at the MIT Media Lab. Without a thorough study of emotions, computer scientists will flounder in programming their robots to mimic and respond to emotions in their interactions with humans and AIs will never connect deeply to humans. Computer scientists, at the very least, need to code the crucially important skill that we might call *artificial empathy*. However, going back to the question posed in the section heading, for a robot to actually be conscious, it would have to both experience time and feel its own emotions. On the basis of the studies reviewed in the last section, there are arguments that suggest that, as early as the world of predators in the Cambrian and ever since, one of the strongest emotions has been the will-to-live, something that integrates emotions with the perception of time. It is an old story, going back to Frankenstein and his monster, retold recently in the science fiction movie *Ex Machina,* that if you are able to create an artificial entity that possesses a will-to-live, it may well go on to threaten his/her human maker. Is allowing for such a threat a necessary component of a conscious robot? In religious terms, we are asking what would lead to the *quickening* of a robot. One can only guess whether opening the Pandora's box of AI will give us such God-like powers. Many would say that even asking this is beyond the realm of science.

BIBLIOGRAPHY

[CC] Cammaerts, M. C. & Cammaerts, R. 2015. Are Ants (Hymenoptera, Formicidae) Capable of Self Recognition? J. of Science, 5, 521-532

[RC] Carnap, Rudof 1999. The Philosophy of Rudolf Carnap. Chicago, Open Court.

[IBC] Cohen, I. Bernard 2016. The Principia. Berkeley, Univ. California Press.

[SD] Dehaene, Stanislas 2014. Consciousness and the Brain: Deciphering how the Brain Codes our Thoghts. New York City: Viking.

[FW] de Waal, Frans 2019. Mama's Last Hug: Animal Emotions and What They Tell Us about Ourselves. New York City, W.W.Norton.

[CD] Darwin, Charles 1872. The Expression of the Emotions in Man and Animals. London, John Murray.

[EP] Eccles, John, & Popper, Karl 1977. The Self and its Brain: the Argument for Interactionism. Abingdon on Thames: Routledge.

[E2] Eccles, John 1990. A unitary hypothesis of mind-brain interaction in the cerebral cortex. Proc. Royal Soc. London B 240, 411-428.

[HG] Gardner, Howard 1983. Frames of Mind: The Theory of Multiple Intelligences. New York City, Basic Books.

[PGS] Godfrey-Smith, Peter 2016. Other Minds: the Octopus, the Sea and the Deep Origins of Consciousness. New York City, HarperCollins.

[HP] Hameroff, Stuart & Penrose, Roger 2014. Consciousness in the universe. Physics of Life Reviews. 11, 39–78.

[SK] Kelley, Sean 2017. Waking up to the Gift of Aliveness. Dec. 25 New York Times Opinion.

[NL] Lane, Nick 2015. The Vital Question. New York City, W.W.Norton.

[MP] McCulloch, W. & Pitts, W. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

[BM] Merker, Bjorn 2007. Consciousness without a cerebral cortex. Behavioral and Brain Sciences, 30, 63-81.

[SM] Montgomery, Sy 2015. The Soul of an Octopus: A Surprising Exploration into the Wonder of Consciousness. New York City, Atria Books.

[RM] Muller, Richard 2016. Now, the Physics of Time. New York City, W.W.Norton.

[DM] Mumford, David 2016. Grammar isn't merely part of language. http://www.dam.brown.edu/people/mumford/blog/2016/grammar.html.

[PB] Panksepp, Jaak & Biven, Lucy 2012. The Archeology of Mind: Neuroevolutionary Origins of Human Emotions. New York City, W.W.Norton.

[WP] Penfield, W. & Jasper, H. 1954. Epilepsy and the Functional Anatomy of the Human Brain. Boston: Little, Brown and Co.

[AR] Rosenberg, Alex 2016. Why you don't know your own mind. July 18 New York Times Opinion.

[ET] Tolle, Eckhart 2003. The Power of Now. Vancouver, Nameste Publishing.

[PW] Werbos, Paul 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University.