David Mumford

Archive for Reprints, Notes, Talks, and Blog

Professor Emeritus
Brown and Harvard Universities
David_Mumford@brown.edu

Can an artificial intelligence machine be conscious?

April 11, 2019

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 machine 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singularity," a point in time when super-AIs will create a wholly new world that leads to the extinction of the now superseded human race (predicted by some to happen around 2050).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 when the first computer attained 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 too. Who has the expertise to work out how all this impacts our understanding of

consciousness? And even talking about the relevance of religion to any scientific advance is anathema to today's intelligencia. 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 seem to live in separate universes.

1. What's missing in state-of-the-art AI?

Let me start by making some 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. More importantly, in 1974, Paul Werbos wrote a PhD thesis 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 my previous blog post entitled "Grammar isn't merely part of language", 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 is we can synthesize new data. 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 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 human intelligence until this is solved.

If the artificial intelligence is to demonstrate human intelligence, we

had better define what human intelligence *really consists in*. 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. Or does intelligence mean solving "Jeopardy" questions?; remembering more details about more events in your life?; composing or painting more skillful works? Yes, sure it might, but wait a minute: 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. Computer scientists have indeed looked at the need to model other "agents" knowledge and plans. A well known example 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 A must send another message to B that he did get the earlier message to be sure B will act. More messages are needed (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. This must include knowing what they themselves know and what they don't know. But arguably, this is all do-able with contemporary code.

2. We need Emotions #$@*&!

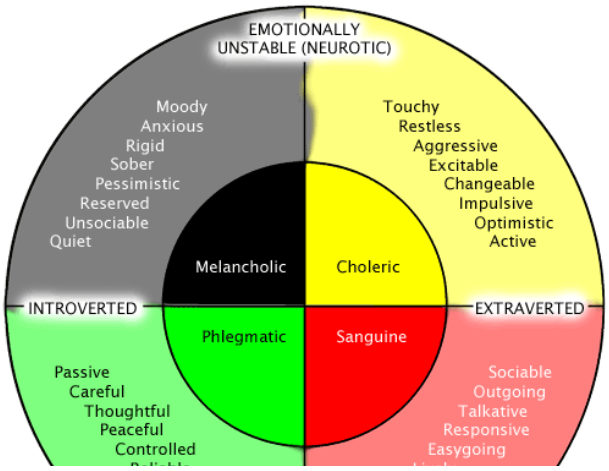
However, in this game-theoretic world, an essential ingredient of human thought is missing: emotions. Without this, you'll never really connect to humans. I find it strange that, to my knowledge, only one computer scientist has endeavored to model emotions, namely Rosalind Picard at the MIT Media Lab. Even 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, says, with regard to both human and animal emotions:

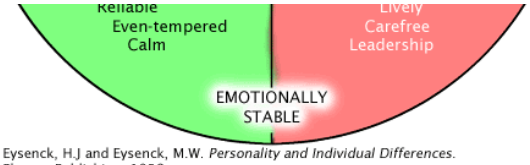
> We name a couple of emotions, describe their expression and document the circumstances under which they arise but we lack a framework define them and explore what good they do.

(Is this possibly the result of the fact that so many who go into science and math are on the autistic spectrum?) One psychologist clearly pinpointed the role emotions play in human intelligence. Howard Gardner's classic book *Frames of Mind: The Theory of Multiple Intelligences* introduces, among a variety of skills, "interpersonal intelligence" (chiefly understanding others' emotions) and "intrapersonal intelligence" (understanding your own). 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 approach to classifying 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), 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/

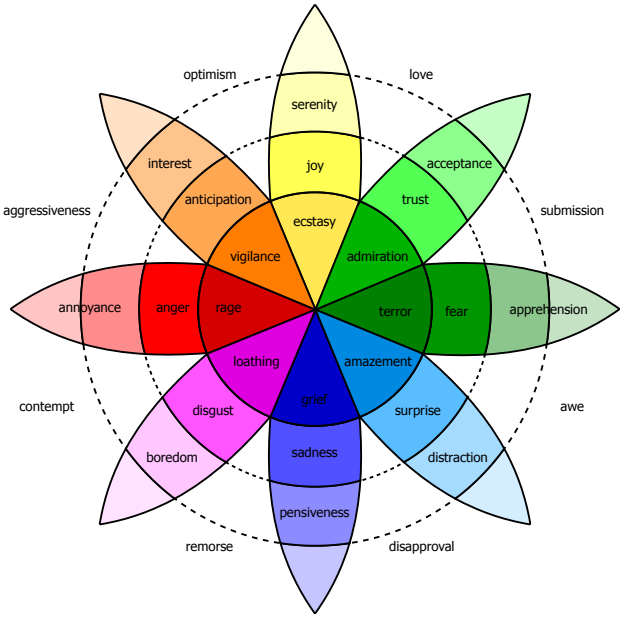




Eysenck, H.J and Eysenck, M.W. *Personality and Individual Differences.* Plenum Publishing, 1958.

phlegmatic being relaxed, choleric/melancholic always seeking more. This classification was developed in recent times by Hans Eysenck, whose colorful version is here:

The modern study of emotions goes back to Darwin's book *The Expression of the Emotions in Man and Animals* 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: There really is an open ended list of secondary emotions, e.g. shame, guilt, gratitude, forgiveness, revenge, pride, envy, trust, hope, regret, loneliness, frustration, excitement, embarassment, disappointment, etc., etc. which don't seem to be just blends but rather grafts of emotions onto social situations with multiple agents and factors intertwined. Frans de Waal in his book (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, e.g. in his and Lucy Biven's book *The Archeology of Mind: Neuroevolutionary Origins of Human Emotions*. 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. Their list is quite different from Darwin's though partially overlapping. They identify 7 primary affective states:

1. seeking/exploring
2. angry
3. fearful/anxious
4. caring/loving
5. sad/distressed
6. playing/joyful
7. lusting

An aside: I am not clear why he does not add an 8th affective state: *pain*. Although not usually termed an emotion, it is certainly an affective state of mind with sub-cortical roots, a uniquely nasty feeling and something triggering specific behaviors as well as causing specific facial expressions and bodily reactions. They go further in Chapter 11 to propose 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)) coordinates all the above affective states and gives rise to what they call *core self* or consciousness.

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, 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. All humans seem to recognize nearly the same primary and secondary emotions when they occur in our friends and artificial intelligences will need to be able to do this too. Without this analysis, computer scientists will flounder in programming their robots to mimic and respond to emotions in their interactions with humans, in other words to possess the crucially important skill that we should call *artificial empathy*. I would go further and submit that if we wish an AI to actually possess consciousness, I believe it must, in some way, have emotions itself. A good way to probe more deeply at the link between consciousness and emotions is to look at non-human animals and see what we know.

3. Consciousness in animals

I want to submit that if we seek to guess whether AI's can acquire consciousness, we should first ask whether animals have consciousness. Let me start by saying to whatever person may be reading this blog post: *I believe that you, my friend, 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, 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 both emotions and 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 referred to above and Jaak Panksepp's book (op.cit.) 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 the same reactions as human babies when their bellies are tickled (see p.367, op. cit.). For me, these books and many others and, of course, my own meagre experiences with owning dogs, pigs, horses and with watching 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* 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, *The Vital Question*). 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). Although I personally prefer here to be agnostic, it is perfectly possible that a mite of consciousness resides in these cells.

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 but I'm not aware of any 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 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 all growing shells for protection, all this in the Cambrian age 540-485 mya. 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 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. (See Godfrey-Smith's book, 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. Damn. Another approach 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, from M-C Cammaerts and R. Cammaerts, J. of Science, v. 5, 2015, pp.521-532).

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 how humans use their cerebral cortex in a supervisory role, 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.

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, Versed is an anesthetic that creates a half conscious/half unconscious state. 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 is pushing it backwards, i.e. they anticipated the push back (Godfrey-Smith, 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.

4. The experience of time and consciousness

I want to go back to the initial question of whether AIs can have consciousness. Much of this last section will annoy many people who read it: I need to cross a further bridge and talk about things that are usually classed not merely as philosophical but as religious or spiritual. I do not want to be "terminally agnostic". Religion has been characteristic of human society since its earliest beginnings and, except for the occasional atheist, has been a central part of everyone's life until some point in the 20th century. With the rise of modern medicine, doctors have replaced ministers as the principle go-to persons when illness strikes and, as I said above, today's intelligencia pays little attention to religion. But I cannot respect rabid atheists like Richard Dawkins who disrespect this entire history and the people who lived it.

I want to start off by repeating a point I made in my previous blog "Let the mystery be". Although sentience, that is sensing the world and acting in response to these sensations, together with the corresponding brain activity, is often considered an essential feature of consciousness, I don't believe that. 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 all 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). I can't say I have experienced this, though I tried a bit. But it makes sense to me because of some times when I started on this path and found some measure of mental peace and quiet. I propose instead that the experience of the flow of time is the true core of consciousness, somewhat in the vein of Eckhart Tolle's "The Power of Now". It rests on the idea that experiencing 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.

I want to quote from the two most famous physicists in order to amplify this idea. Firstly, Newton, in his *Principia* states:

> 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 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. He was, however, fully aware that people experience what Newton was describing and wondered if this, with its notion of the present, could have a place in physics. Though he never wrote about this, he had a conversation with Rudolf Carnap in which he made this point. (My thanks to Steven Weinstein for telling me about this conversation.) Here is how Carnap 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.

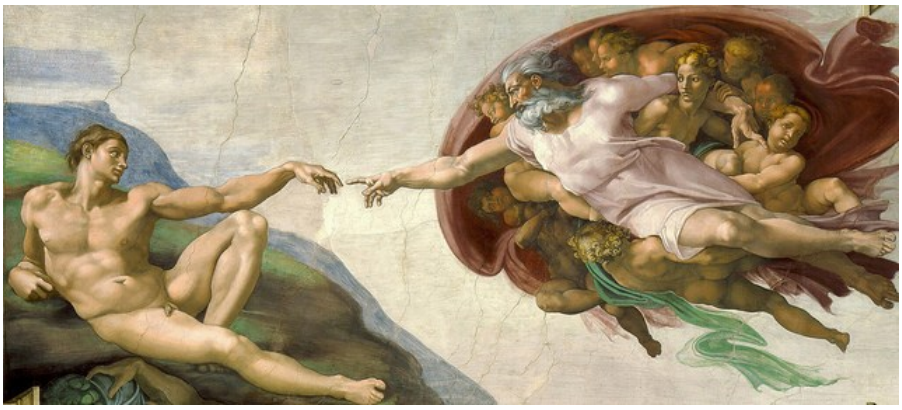
Yes, yes, yes, that's what I'm talking about! How wonderful to hear it from Einstein.

Where does this leave our discussion in the last section? I don't want to suggest that sentience has nothing to do with consciousness. I think the two are highly correlated and that Buddhist monks perform a sort of

mental gymnastics. I want to list the properties that I have argued for, properties that circumscribe to some extent what consciousness is:

> 1. Consciousness is a reality that comes to many living creatures sometime around birth and leaves them when they die, creating a feeling of "moving" from past to future along a path in space-time as well as feeling sensations, emotions and their body movements.
> 2. Consciousness has degrees, varying from utterly vivid (e.g. positive feelings like love and negative feelings like pain) to marginal awareness. The brain has, moreover, an unconscious as well as a conscious part, activities, even thoughts, that do not reach consciousness..
> 3. Consciousness occurs in many creatures including, for instance, octopuses as well as mankind.
> 4. Consciousness gives us the belief that we have *free will*, that we can make choices that change the world. This has some relationship to quantum mechanics.
> 5. Consciousness is not describable by science, it is a reality on a different plane.

Points 1. and 5. follow Einstein's comments quoted above, points 2. and 3. are the ideas from the previous section (and from Dehaene's book *Consciousness and the Brain*). Point 4. was one of the main issues discussed in my earlier blog and on which I hope to write later (see e.g. Stapp's book *Mindful Universe*). I think it fair to say that religions are unanimous is embracing points 1. and 5. and saying that consciousness, at least in humans, results from some sort of spirit enlivening, quickening our body, as Michaelangelo depicts it:



My own version of this is to say: consciousness results from "spirit

falling in love with matter". Why love? This is a metaphorical way of expressing the intensity of consciousness and the associated will-to-live that seems universal among animals. "Love" is just the human way of saying the spirit forges an awfully tight bond with matter in which dying is painful. The main point is that if you accept that experiencing time cannot be explained by science, but occurs in definite and not random fashion, then it seems that this experience *has to come from somewhere*. So again, Occam's Razor suggests the simplest path is to use a word proposed by all religions and call it spirit. This is really a minimalist approach, not based on any personal revelation.

So the question with which this post is titled then becomes: what would make a robot seem like a likely place for spirit to want to quicken? Except possibly for pantheists, no one believes that rocks are aware. Everything I've written in this post suggests that the robot somehow had better have authentic emotions, one way or another, for this to happen. Quite a challenge. Amen.

SOME EMAIL RESPONSES

I. My colleague Shiva Shankar at the Chennai Math Institute wrote me: "Yes indeed David, the mysterious ever-present, ever-slipping away Now. Buddhism tries to grapple with it a lot, and Vipassana meditation (which I practice) avers that to be immersed in the Now requires one to have gone beyond the sense of self, and to be suspended or immersed in a very different (and strange) state of existence. (Also) I wanted to ask you about a sentence below, and in your post: when Einstein says the NOW is outside the realm of science, does he mean that you cannot grasp it or its meaning or its implications by any replicable experiment?"
My reply: He means you cannot even define it, science only deals with things that can occur anywhere any time.

II. My colleague C.S.Seshadri wrote me:
"I read your article on AI with great interest. Very interesting to read about consciousness and your skepticism in efforts to build machines endowed with consciousness. In a related manner I have always felt that modern philosophies like Communism don't deal with the emotional issues like jealousy, love, hatred, ambitions etc., which are

central issues for humanity. Only great religions deal with these. I like your quotation from Einstein. With the limitations of human intelligence, life will remain a mystery. Perhaps this cannot be communicated in words and can only be experienced. One cannot discount spiritual experiences. Sometimes I feel spiritual experience is like what we have in mathematics, a sudden flash which puts everything in place."

Thanks Seshadri.

III. My statistics friend Wu, Yingnian at UCLA wrote me:
"Apparently consciousness is about knowing what you see, what you feel, and what you think, and emotion is about what you feel. Thus, an animal may be emotional, but it may not be conscious of its emotion. That is, there seems to be a difference between emotion and consciousness. "I feel emotion may serve as (1) a trigger for action, i.e., a policy. This includes anger and fear, for fight or flight. (2) a surrogate for value (which is ultimately the survival of the genes). This includes pleasure and pain. A lower animal has no capacity for planning. Thus it relies on emotion for taking action. It can be emotional but it does not have consciousness. "For smarter animals (including crow), the emotion does not only trigger action, it often serves as part of the planning computation, i.e., the policy and value play the role of initialization and surrogate in the planning computation, and eventually we act according to planning. However, sometimes the emotion is so strong that it may override planning, for instance, when people become extremely angry (or the situation is extremely urgent that we do not have time to plan), or when people become highly addictive (i.e., the surrogate is completely wrong about the value). "So emotion and feeling about emotion may be two different things. All animals are capable of emotion, but only animals who plan can feel and even express emotion. Sorry I may be completely wrong in this regard. "Planning also seems to require symbolic reasoning (logic and grammar) or at least a very low-dimensional representation and abstraction, in order to make the computation efficient. "

IV. My son Jeremy Mumford at Brown wrote me:
"I got a little confused toward the end, when you talked about Einstein, Buddhist monks, Eckhart Tolle, and the experience of the flow of time

as 'pure consciousness'. Subjectively, the feeling of moving through time does seem to be at the heart of my sense of consciousness. But when we infer consciousness in others (both people and animals), it is not though evidence that they experience the flow of time, but through evidence of reasoning and especially emotion. It's the octopus's playfulness that makes us think it's conscious. So I am not sure why you privilege the flow of time as being 'pure consciousness.' "

Before commenting on these two emails, I feel I need to describe to restate my understanding of consciousness. As I see it, consciousness can be vewed as having three ingredients: a) *cognitive skills*, b) *subjective feelings* like pain and love and c) *the experience of the flow of time*. The first is the basis of what people now study in the "scientific" study of consciousness, especially via human reports of their own thoughts. But I don't believe this is the core because it's easy to imagine your own state without any cognitive activity such as planning, e.g. after a stroke or after trauma leaves you in a disoriented daze. Likewise, feelings come and go and their departure doesn't interrupt your sense of time passing. This leaves the last as the true core of consciousness.

I believe what you are saying, Yingnian, is that there is a divide between lower animals that have emotions for triggering actions and higher animals who are conscious, who plan and who can even express their emotions. This idea seems natural and widespread but my post specifically seeks to cast doubt on this. I don't see why consciousness should rest on the ability to plan as we have conscious awareness in all sorts of situations and states-of-mind having no planning component. I am proposing that emotions are closer to the core of consciousness, that we believe our pets are conscious chiefly because of the emotions they exhibit. This is not a scientific argument but I don't know a scientific reason for being convinced that another human is conscious either.

I agree, Jeremy, that all we have to convince us any animal is conscious are its emotions and intelligent actions. BUT it’s all you have to believe I'm conscious too! On the other hand, the awareness of passing time and being in the NOW feels deeper to me, the core of consciousness that seems absolutely unreachable by experiment. After the advent of relativity, physics ceased to try to model this awareness and it fell into a limbo. There’s no way to convince a solipsist that he/she is wrong as far

as I know. So we must make some sort of myth about spirit.

V. A friend from Florida, John Thomas, wrote me these wide ranging comments:
"Your comment and belief in, “Spirit falling in love with matter, “ echoes Edgar Cayce’s, the American psychic, who said that we are spiritual co-creators, with The Creator, who could manifest and change form, without hindrance. Unfortunately, we became entangled and lost in the material plane and its temptations and delights, particularly sex. He cited Atlantean and Egyptian civilizations as prime examples of the brilliance of creation and the conflicts of ambition, aggression and greed. Hence Miltons “FALL,” and the Bible’s “Eviction from the Garden of Eden.” "Cayce also speculated that there are 9 planes of existence, and the material plane being the lowest, densest and the most difficult. I have read that quantum physicists have speculated there as many as 11 planes of existence. Buddhists, in particular, believe that we experience multiple incarnations until we are enlightened and freed from the bondage of the material plane and return to our spiritual heritage and home. "I was interested to read your, and other respondents’ comments, about “spirit” entering the body, when, of course, there is no proof, or evidence, of its existence, other than the hopeful speculation that we have a “soul,” which is a cop-out on the inevitability of death. “Where there is life there is hope,” suggests that where there is no life i.e. death, there is none. "Speculation on the soul has given way to the much more material consequences of AI. I read Yuval Harari’s, “Homo Deus,' with great interest, after his “Sapiens.” He has no doubt that we are redundant, and will be easily superseded by AI, which, with its rapid transformations, is already surpassing our best intelligences as its algorithms compound and create new advances that are beyond our abilities, and in some cases, comprehension. This AI advantage is expanding at a rapid rate to the degree that machines will program their own progressions, with unimaginable consequences. "We take great pride in our emotions, and human feelings. Yes, they give us love and tenderness, but hate and aggression as well, as our bloody history shows. I remember reading in the 60’s, Robert Audrey’s “The Territorial Imperative,” and Konrad Lorenz’s, “On Aggression,” which were profound and challenging. So, of course, was Marshall McLuen’s “The Medium is the Message.” All were prescient and have come bear on the most immediate issues of our world to-day. You could include

Rachel Carson's, "Silent Spring," to complete the picture.
"Consciousness has never been fully understood, but there are myriad speculations. Buddha proclaimed all is, "Maya," an illusion, and excessive attachment to material gain, which describes our entanglement in the material plane. "Harari dimisses religion as myth, superstition and a means to subordinate, and dominate, groups to some creed and belief system. It is inevitably at war with others, as is increasingly evident in our world to-day, as we revert to tribal behavior.
"Whether AI takes over, or not, the human race is in great peril, and, even more so, the planet. Malthus, in the 18th century, predicted that the human race would eventually infest and overwhelm the planet and its food supply given its continued rate of reproduction. Unfortunately, those who can least afford to have and raise children are breeding at the fastest rate. The Catholic Church historically encouraged unrestricted reproduction, and the Muslims are breeding at the fastest rate to-day. It is unsustainable, but short of legal control (the Chinese tried, and abandoned, the "one child" policy) there is no solution to hand, other than war, disease, or AI interference. AI has an advantage as their diet is electricity, and one assumes it will manage to organize clean energy.
"I intended to post this on your blog, but was unable to process."