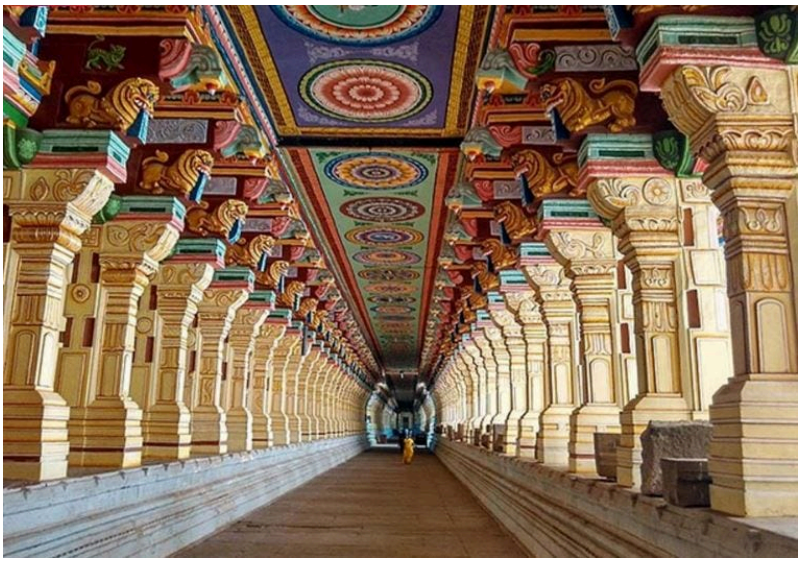
Artificial Intelligence, Consciousness and the Self

1.

By Subhash Kak

Dec 07, 2021 12:43 AM · 11 min. read · View original



The Infinite Corridor of the Ramanathaswamy Temple: Wikimedia Commons

1.

There are many scientists and engineers who believe that computers will eventually become conscious. Computer scientists Bernard Baars and Stan Franklin wrote in 2009: “consciousness may be produced by … algorithms running on the machine.” MIT Technology Review in October 2017 opined that this may happen in the not-too-distant future. Given the fact that biological evolution is slower than technological evolution, there is fear that humans will be unable to compete with sentient machines. No wonder many prominent voices in the scientific and the tech worlds are claiming that artificial intelligence is leading humanity to a catastrophe. Stephen Hawking told the BBC in 2014, “I think the development of full artificial intelligence could spell the end of the human race.” Speaking to the National Governors Association in 2017, Tesla CEO Elon Musk said that “AI technology is a fundamental risk to the existence of human civilization.”

So what is going on? Can the “ghost of the self” in the body-machine be fully explained in terms of computation? Is it

indeed inevitable that machines will become conscious unless AI is strictly regulated? Some futurists and sci-fi writers are imagining a marriage of sorts between AI and brains so that, someday, technology will make it possible for humans to become "posthuman," transcending the limits of the human condition. There are others who believe that the only way to make sense of all the scientific facts is to take reality as a simulation, an idea that was used by Hollywood in movies, such as The Matrix and Watch Gamer. Another scenario is to imagine that once humans learn how to completely characterize brains, they will be able to copy themselves into computers, creating their emulations, or *ems*, in the process. Such *ems* could be replicated easily and so they will quickly outnumber real people.

In any event, if and when artificial intelligence evolves into conscious machines, that will surely mean the end of the world as we know it. Conscious machines will most likely take over the world and enslave humans, if not worse. Given this possibility, the prospect of machine consciousness is a most

pressing challenge facing humanity at this time.

Some may quibble over the idea of “consciousness” in machines. If we agree to use it in the sense of “awareness” and “personhood”, machines with it will have access to greater autonomy than machines that lack it and, therefore, they would potentially replace humans in most jobs and raise insoluble problems of ethics and morality. Even if each human is provided a universal basic income for food, shelter and entertainment, as is being proposed by many economists — and the machines leave us alone — it will be a dystopic life.

There are many arguments in support of the idea of an eventual evolution of consciousness in computers. Since we cannot deny that the brain is a machine, then other machines with appropriate architecture should also achieve consciousness. Some object and say that brains, unlike classical computers, perform quantum operations deep in the neural circuitry. However, one can postulate future quantum computers that

mimic brain behavior in all possible ways.

Now there are cognitive tasks during which the agent does not have a sense of subjective awareness. Such cognitive tasks are also being done by the machine, and it does so faster and with more reliability.

Awareness is something beyond computation for it is the ability to halt the processing in the brain machine at will to take stock of what is going on. Such a halting problem was considered by Alan Turing in 1936 and he showed that no algorithm can determine whether given a description of an arbitrary computer program and an input, the program will halt or continue to run forever.

If halting to an arbitrary input at randomly chosen time is impossible from a computability point of view, and the aware mind does it, then one may conclude that *consciousness is not computable.*

2.

In our intuition, consciousness is a category that is dual to physical reality. We apprehend reality in our mind and not in terms of space, time and matter. This experience varies based on brain states and one may conclude that non-humans experience it differently from us. It is also significant that in our conscious experience we are always outside of the physical world and witness ourselves as apart from our bodies. Even in scientific theory, as for example in classical mechanics, the observer is away from the system, even though there is no explanation of the observer within the theory.

To help answer the question whether machines will become conscious we must go back to the question of the nature of reality. Is the world a machine described by its parts and their interconnections, or is it fundamentally knowledge? The first view is called *ontic* (from ontological, that is related to structure), and the second is called *epistemic* (from epistemological, related to knowledge). In philosophy, these are the positions of two different schools, one believing that reality is *being*, and

the other that it is *becoming*. The conception of the world as *being* is associated with materialism, while that of *becoming* assigns a more significant role to the observers.

3.

Last year, I was part of a series of weeklong workshops organized by SRI International, Menlo Park in different locations in the United States and Cambridge, UK to consider this question of whether machines at some future time will become conscious. The thirty-odd participants in these workshops included computer scientists, physicists, neuroscientists, philosophers, and strategic thinkers.

We took stock of the many difficulties with the conception of selfhood in a machine paradigm. Standard neuroscience accepts the doctrine of the identity of brain and mind. In this view, mind emerges from the complexity of the interconnections and its behavior must be completely described by the corresponding brain function leaving no room for agency of the individual. But no specific neural correlate of

consciousness has been found and consciousness cannot be localized.

The selfhood of humans leads to paradoxes related to autonomy and freedom. Humans reject the idea that they are mere machines, yet they often equate their “self” with the machinery of the body. On the other hand, the human’s self-image is that of the body, together with transient thoughts, which is overseen by an observing “I” within.

We make a distinction between the “autobiographical self” related to one’s memories and relationships, and the “core self”, which is rooted in the momentary present. The “autobiographical self” is partly the result of one’s imagination since it is an interpretation of the past and it includes hopes for the future. The “core self” is elusive; it is the light that shines on things around and associates with them in time and space.

The experiments of Benjamin Libet showed how decisions made by a subject arise first on a subconscious level and only afterward are translated into the conscious decision. Upon a retrospective

view of the event, the subject arrives at the belief that the decision occurred at the behest of his will. In Libet's experiment the subject was to choose a random moment to flick the wrist while the associated activity in the motor cortex was measured. Libet found that the unconscious brain activity leading up to the conscious decision by the subject began approximately half a second before the subject consciously felt that he had taken his decision. But this is not to be taken as an example of retrocausation; rather, this represents a lag in the operation of the conscious mind in which this construction of reality by the mind occurs.

The participants at the Workshops agreed that AI-machines of the future will be able to emulate all cognitive tasks and by implication able to replace humans at all kinds of jobs. But they were split on whether machines will be conscious like humans are. The split turned out to be based on the idea that the phenomenon of consciousness could come in two different varieties which I call little-C and big-C. If all there is to consciousness is little-C then machines

will be conscious. But if human consciousness is actually big-C, then machines will fall short.

4.

So what are these two conceptions of consciousness? Little-C is consciousness emerging out of the complexity of the brain processes. It is emergent in the same sense that biology is emergent on chemistry, which, in turn, is emergent on physics. It is similar to the thought, generally ascribed to Buddhism, that consciousness arises on ground that is emptiness (*śūnyatā* in Sanskrit). In this view, if a sufficiently complex machine is able to emulate the processes in the brain, it will be conscious.

On the other hand, big-C assumes consciousness is something that is apart from the physical reality. Philosophically, this is the position of Vedanta, in which the mental and the physical phenomena are two aspects of the same reality, like two sides of a coin. (As an aside, the Buddha declared on his deathbed that he agreed with the Vedic position.)

The pioneers of quantum theory used the so-called Orthodox Copenhagen Interpretation (CI) to understand the mathematical formalism of the theory, where the underlying idea is of big-C. CI assumes complementarity at different levels and this includes the duality of matter and mind or object and subject.

One may devise scientific experiments on creativity to further investigate the two views of consciousness. It appears that the creative moment is not at the end of a deliberate computation. There are many autobiographical accounts of dreams or visions that preceded specific acts of creativity. Two famous examples of this are Elias Howe's 1845 design of the modern sewing machine, and August Kekulé's discovery of the structure of benzene in 1862.

The life of the self-taught Indian mathematician Srinivasa Ramanujan, who died in 1920 at the age of 32, is evidence in favor of big-C consciousness His long-forgotten notebook, which was published in 1988, contains several thousand formulas that were well ahead of their time, without explanation of how

he had arrived at them. When he was alive, he claimed that formulas were revealed to him in his asleep.

But how might matter and mind mutually influence each other? One possibility is through the act of observation which causes the collapse of the state function in quantum theory. If the observation is made repeatedly, the system state will freeze. Called the Quantum Zeno Effect, it has been demonstrated in the laboratory.

5.

Even if one were to dismiss accounts of creativity as nothing but coincidence, the ontic understanding of reality becomes problematic when one brings in information into the mix, as is done extensively in modern physics. This is because information implies the existence of a mind, which category lies outside of the realm of physics.

Information or entropy cannot be reduced to local operations by any reductionist program. It requires the use of signs derived from global properties and the capacity to make choices which,

in turn, implies agency. Entropy is a measure of disorder and in certain situations may be measured by temperature. This is how after determining the many different states associated with a black hole, Stephen Hawking was able to postulate a corresponding temperature and speak of radiation from it. But temperature like entropy or information cannot be associated with a single particle.

Information in a communication involves two things: first, commonalities in the vocabulary of communication between the two parties; and second, the capacity to make choices. The common vocabulary requires that the underlying abstract signs used by the parties be shared, which stresses the social aspects of communication.

Schrödinger, one of the creators of quantum theory, stressed the epistemic nature of the state function. He was emphatic that “Consciousness cannot be accounted for in physical terms. For consciousness is absolutely fundamental. It cannot be accounted for in terms of anything else.”

6.

The idea of consciousness requires not only an awareness of things but also the awareness that one is aware. If awareness is some kind of a measurement, it should have a reference. This, in turn, poses two problems: first, what is the reference for awareness; and, second, how does consciousness choose between various possibilities?

The problem of the referent in awareness is an old one. If we postulate a single universal, transcendental consciousness, the individual's empirical consciousness is a projection and the referent for it must be the universal. In Vedanta, the analogy of the same sun reflecting in a million different pots of water as little suns is provided to explain the empirical consciousness of the individual. This is also similar to the view of Plato who invokes an abject in a cave that cannot be seen directly whose shadows on the wall are accessible.

In Western philosophy, René Descartes proposed that consciousness resides within an immaterial domain called *res cogitans* (the realm of thought), to be

contrasted from the domain of material things, called *res extensa* (the realm of extension), and he assumed that the two realms interact in the brain, but this Cartesian dualist position is no longer taken seriously. On the other hand, Immanuel Kant arrived at a resolution similar to that of the Vedic tradition by arguing that empirical consciousness must have a necessary reference to a transcendental consciousness (a consciousness that precedes all particular experience). The universal or transcendental position is generally unacceptable to mainstream scientists who insist on reductionist models.

William James spoke of two kinds of selves: the self as knower (the "I"), and the self as known (the "me"). Each person's self is partly subjective (as knower) and partly objective (as known). The objective self itself may be described in its three aspects: the material self, the social self, and the spiritual self. Narrative self-reference is in contrast to the immediate, knowing "I" that supports the notion of momentary experience as an expression of selfhood.

James believed that as knower, the self is comprised of different mental states. Thought has no constant elements and every perception is relative and contextualized. States of mind are never repeated, and whereas objects might be constant and discrete, thought is constantly changing and mental states arise out of choices that are made by the mind. James believed that thought flows, and thus he could speak of a stream of consciousness.

If one were to find the boundaries between the “me” and the “I” of consciousness, it becomes essential to find a “minimal” sense of self. It is easy to speak of the intuition that there is a basic or primitive something that is the true self, and much harder to provide evidence for such belief. Conceptually, there must be something permanent — a bed rock -– underlying the stream of consciousness.

To deal with the empirical pre-conscious or conscious awareness, one can postulate a hierarchical model of consciousness with independent and distributed neural structures at the

lowest level. The speed of the binding of the attributes would depend on the complexity of the communications and the relationships between the modules. Interference between various levels and the tangled nature of the information flow can help explain many illusions of perception.

7.

Memory is one element that leads to the nature of the corresponding consciousness state. The mind must select from the pool of memories and this selection may not be made consciously and it may be determined by the stream of previous consciousness states and the emotional state of the subject. It is to be expected that the executive control processes play an important role in the selection.

Furthermore, repeated selection of certain memories at the expense of others may affect the recall process, causing unwanted memories to be pushed back into the unconscious. Mechanisms can be recruited that prevent unwanted declarative memories from entering awareness, and that this

cognitive act has enduring consequences for the rejected memories.

In the state called mindfulness, the executive control appears to be able to recruit memories with great ease. Some see mindfulness as meta-awareness (self-awareness) with ability to effectively modulate behavior (self-regulation), and a positive relationship between the self and others.

Although awareness by itself is an all or nothing phenomenon, the state of consciousness depends on the degree to which preconscious and memory states are accessible to awareness. The uncoupling of perception from sensory inputs supports the idea of the disembodied consciousness state. A consciousness state that is decoupled from specific memories of the individual indicates that such states may be fundamental to reality and do form a part of the ontological reality.

Many aspects of reality in the fields of physics, mathematics, and brain states either have paradoxical aspects or are not computable. Therefore, to assume that machines based on logic and

mathematics can emulate all natural systems fully is incorrect. To put it differently, the reality described by machines is of a kind different from that of natural systems.

But this does not mean that AI machines will not displace humans from most jobs. And even if machines did not become conscious, there will be increasing tendency on the part of humans to treat them *as if* they were conscious.