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, part II?

July 12, 2019

Since posting "Can an artificial intelligence machine be conscious", I have read two books which argue for the belief that conscious artificial intelligences are on their way and we should prepare ourselves. I want to discuss both books and evaluate my own arguments in their light. My first read was fiction: Ian McEwan's latest book *Machines Like Me* in which a small group of conscious robotic men and women are manufactured and sold around the world. His novel is arguably an ideal way to make the concept of a conscious robots both plausible and frightening. The two chief protagonists have no doubt that their Adam (as he is named) is conscious nor does his character Turing (a version of Turing who lives a long and amazing scientific life). But it does not end well! My second read was *The Ego Tunnel*, by the German philosopher Thomas Metzinger. This book exhaustively examines what consciousness is from biological, psychological, information-theoretic and philosophical perspectives. It presents very relevant data from Out-of-Body Experiences, lucid dreaming and much else. After an analysis of what is going on in human brains, he writes a section entitled "How to build an artificial conscious subject and why we shouldn't do it" outlining how it might indeed be done. Let me look at both books in more detail.

The Novelist

Spoiler alert: I can't discuss *Machines Like Me* without revealing the plot. McEwan plunges right in with their robot Adam falling in love and sleeping with Miranda, the girlfriend of the Adam's purchaser Charlie. Although needing to be regularly recharged by a plug in his navel, he has been loaded with basic human emotions, partly by Charlie and Miranda clicking a set of online choices. Next he breaks Charlie's wrist when Charlie inadvisedly reaches for the off button on his neck. But they soldier on when Adam apologizes to Charlie, only to find in the denouement that his idea of moral behavior is totally out of synch with humankind's waffling moral compromises, with actions that send Miranda to jail. Charlie, out of his love for Miranda, smashes in Adam's skull and Turing brands him a murderer.

McEwan certainly makes hay from my precise point: that human emotions are extremely complex and convoluted and thus one has to question whether a robot can ever truly "understand" them. Yet I have argued that an essential part of being conscious is precisely "feeling" emotions. I put this in quotes as feeling and understanding are words that touch on what consciousness is. It seems to me that McEwan is making too fine a point by allowing Adam many intense emotions yet failing to give him any deeper understanding of how emotions work.

His failure highlights the human behavior pattern expressed by the word "loyal". This word refers to a mix of emotions and of patterns of actions, both past and future and is typical of the complex interweaving of emotions and activities in human beings. For instance, the central principles of Scottish ethics might well be thrift, honesty and loyalty, all three being emotionally freighted activities. Adam is thrifty and honest but fails on the demands of loyalty. On the other hand, my cousin Ruth Silcock wrote a series of children's books about a cat named Albert John. In her first book, she wrote "Albert John was a loyal cat", *assuming that this concept was perfectly clear to her young readers.* But not so for Adam. Thus, by and large, McEwan is agreeing with my belief that modeling human emotions and their resultant activities in a robot is a huge hurdle, even though his characters do see their robot as emotional enough to be deemed conscious.

The Philosopher

Metzinger's book is easily the most insightful dissection of the nature of consciousness that I have read. His basic thesis is that our brains construct for us a *phenomenal self-model*, by which he means "the conscious model of the organism as a whole that is activated by the brain" and that he also calls the Ego (p. 4). He elaborates this as follows (p.7):

> First our brains generate a world simulation, so perfect we don't recognize it as an image in our minds. Then they generate an inner image of ourselves as a whole. This image includes not only our body and our psychological states but also our relationship to the past and the future as well as to other human beings. The internal image of the person-as-a-whole is the phenomenal Ego, the "I" or "self" as it appears in conscious experience.

He says we feel we are consciously *having* the experiences that our bodies encounter in the world because this integrated inner image of ourselves is firmly anchored in our feelings and bodily sensations and because we are unable to recognize our self-models *to be just models*, because they are *transparent* like a glass window through which we see the world. Thus he is led to describe the life we lead as an *Ego Tunnel*. Our minds are filled by a model that we take for reality, hence we are in a tunnel through which we move as time goes on. Although he does not mention Schopenhauer, much of this theory seems similar to Schopenhauer's ideas: *Die Welt ist meine Vorstellung* (The world is my representation) is the assertion with which he opens his magnum opus *Die Welt als Wille und Vorstellung*.

Metzinger makes a great deal of the so-called rubber hand illusion. Here, the subject sits at a table with his left hand behind a barrier, but a rubber left hand is placed on the table in front of him. Then the rubber hand is tickled by a feather while, invisibly, his real left hand is also tickled. After a certain amount of time, the subject begins to feel the rubber hand is his own, that an invisible arm connects it to his body and tickling it alone causes him to feel his real hand is tickled. Metzinger interprets this as tricking the mind into altering its self-model into an unreal representation that still feels totally real. Similarly, he discusses at length phenomena like out-of-body

experiences and lucid dreams (where you are aware you are dreaming but still feeling you are living a vivid convincing dream world). Oddly, he doesn't describe some of the other virtual reality experiments like the one where, wearing goggles that show you walking over a virtual cliff, you fall down with genuine fear (though actually onto a carpet in an empty room). I was a subject and experienced this at Brown. Nor does he discuss the vast virtual world in the movie "The Matrix" and the present vogue for virtual reality goggles and immersive entertainment. But surely these only reinforce his argument that we live in a self-model and can all too easily be tricked into taking an alternate world as reality.

The Sage

My favorite story from the rich legacy of Hindu Mythology is the story of the sage Narada and his quest to understand Vishnu's *Maya*. It illustrates that Metzinger's phenomenal self-model has antecedants that go back at least to the first millenium BCE. It starts with Narada performing so many austerities that he acquires the spiritual power to ask Vishnu for a boon. He asks for an understanding of Maya (an ancient sanskrit word for "illusion"). The story goes on, in the telling by Heinrich Zimmer (*Myths and Symbols in Indian Art and Civilization*, pp.32-34):

> "Show me the magic power of your Maya ," Narada had prayed, and the God replied, "I will. Come with me;" with an ambiguous smile on his beautiful curved lips. From the pleasant shadow of the sheltering hermit grove, Vishnu conducted Narada across a bare stretch of land which blazed like metal under the merciless glow of a scorching sun. The two were soon very thirsty. At some distance, in the glaring light, they perceived the thatched roofs of a tiny hamlet. Vishnu asked: "Will you go over there and fetch me some water?" "Certainly, O Lord," the saint replied, and he made off to the distant group of huts. The god relaxed under the shade of a cliff, to await his return.
>
> When Narada reached the hamlet, he knocked at the first door. A beautiful maiden opened to him and the holy man experienced something of which he had never up to that time dreamed: the enchantment of her eyes. They resembled those of his divine Lord

and friend. He stood and gazed. He simply forgot what he had come for. The girl, gentle and candid, bade him welcome. Her voice was a golden noose about his neck. As moving in a vision, he entered the door. The occupants of the house were full of respect for him, yet not the least bit shy. He was honorably received, as a holy man, yet somehow not as a stranger; rather, as an old and venerable acquaintance who had been a long time away. Narada remained with them impressed by the cheerful and noble bearing, and feeling entirely at home. Nobody asked him what he had come for; he seemed to have belonged to the family from time immemorial. And after a certain period, he asked the father for permission to marry the girl, which was no more than everyone in the house had been expecting. He became a member of the family and shared with them the age-old burdens and simple delights of a peasant household.

Twelve years passed; he had three children. When his father-in-law died he became head of the household, inheriting the estate and managing it, tending the cattle and cultivating the fields. The twelfth year, the rainy season was extraordinarily violent; the streams swelled, torrents poured down the hills, and the little village was inundated by a sudden flood. In the night, the straw huts and cattle were carried away and everybody fled. With one hand supporting his wife, with the other leading two of his children, and bearing the smallest on his shoulder, Narada set forth hastily. Forging a head through the pitch darkness and lashed by the rain, he waded through slippery mud, staggered through whirling waters. The burden was more than he could manage with the current heavily dragging at his legs. Once, when he stumbled, the child slipped from his shoulder and disappeared in the roaring night. With a desperate cry, Narada let go the older children to catch at the smallest, but was too late. Meanwhile the flood swiftly carried off the other two, and even before he could realize the disaster, ripped from his side his wife, swept his own feet from under him and flung him headlong in the torrent like a log. Unconscious, Narada was stranded eventually on a little cliff. When he returned to consciousness, he opened his eyes upon a vast sheet of muddy water. He could only weep.

"Child!" He heard a familiar voice, which nearly stopped his heart.

> "Where is the water you went to fetch for me? I have been waiting more than half an hour." Narada turned around. Instead of water he beheld the brilliant desert in the midday sun. He found the god standing at his shoulder. The cruel curves of the fascinating mouth, still smiling, part with the gentle question: "Do you comprehend now the secret of my Maya?

In Metzinger's language, I would interpret this story as follows: Vishnu put a fork in Narada's Ego Tunnel and led him down the new fork by his request for water. The new fork was long and ultimately led to Narada experiencing his own drowning. But then Vishnu made the new fork rejoin the old with a touch of a cruel smile on his face. Thus Maya can be seen as a description of one's phenomenal self-image, a convincing reality but only a small window into what is out there, constructed by our limited consciousness.

Time

In a later Chapter "A Tour of the Tunnel", Metzinger discusses the various challenges to how the self-model comes to be so real as to create a sense of consciousness. It is here that he parts company with me, in the section entitled "The Now Problem: A Lived Moment Emerges" (p.34). I couldn't agree more with his sentence "A complete scientific description of the physical universe would not contain the information as to what time is 'now' ". As I have written, science, almost by definition, seeks laws that hold regardless of time and place and leaves anything to do with past/present/future to historians and futurists. As quoted in my previous post, Einstein himself felt 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". One would assume that Einstein's thinking was also motivated by his special relativity in which simultaneity was shown to have no physical meaning, so that one cannot think of the universe as a whole having a past and a present, ever unrolling as future events come to pass. Because of relativity, each person has his or her own Now and, once space travel becomes real, it will be clearly impossible to believe in an objective Now. In Narada's story, 12 years for Narada is half an hour for Vishnu and space travel can make that happen for two human beings.

And, indeed, Metzinger goes on to state "My idea is that this simultaneity is precisely why we need the conscious Now". It is well known that the mind plays fast and loose with simultaneity, so that two signals may be perceived consciously as occurring in the opposite order to their occurrence in the physical world. Temporal order seems to be, to some extent, a construction the mind makes as best it can. But now Metzinger reverses the logic. From the implication that experiencing a Now implies experiencing simultaneity, he wants to say that experiencing simultaneity creates the experience of the Now. He argues that creating a common temporal frame of reference for all the mechanisms in the brain leads to the inner model of the world around such a Now (p.36). Here I cannot follow him: all computers have a clock and organize their computations and communications accordingly. But I don't think they have consciousness. Neurons transmit signals much slower than light and this probably has something to do with the temporal order mistakes the brain makes. But getting most things temporally straight doesn't seem to me enough to produce the conscious sense of Now. You, the reader, will be reading this at some point in the future and I will then be doing something else (or will be dead), thus each of us is living through our own unique sequence of Nows. We may be somewhat in synch when we are reading the same issue of the New York Times but nonetheless our Nows are distinct and our synchronization is doomed to have some degree of uncertainty.

As mentioned above, Metzinger spells out the application of these ideas to the construction of conscious robots in the later section"How to build an artificial conscious subject and why we shouldn't do it" (p.190). On p.192, he describes the construction in four steps. The first is to endow the machine with a continuously updated integrated inner image of the world. The second is to organize its internal information flow temporally, resulting in a psychological moment, an experiential Now. The third is to be sure that these internal structures cannot be recognized by the artificial Conscious system as internally generated images, so they are transparent. The fourth step is to integrate an equally transparent internal image of itself into the phenomenal reality. None of these seem to Metzinger or me be considered to be impossibly difficult. But it's the second where I feel he is assuming too much happens as a result of the silicon activity. The Now seems to me the truly magical step, the step that creates what Popper calls world II and

others call spiritual.

The fact that our lives are lived as a trip down the river of time, that we are always conscious of being at a specific place surrounded by a local bit of the 3 dimensional world which changes as "time goes on", all this seems obvious and commonsensical, not magical at all. But this is because this experience of time is the core of everyone's consciousness, everyone's daily lives, not because in physics or in any other science is there anything like the flow of time with a present instant, lit up like a lighthouse. In physics, time is static, simply one way to put coordinates on the 4-dimensional panoply of all events, past,present and future and in any place whatsoever. We can artificially construct a mathematical "flow", a one-dimensional group of homeomorphisms of space-time, but no such flow is given by physics and it has no distinguished set of points called the present moment. To live in a world of time seems to me a wonderful gift and I have no clue how, like God and Adam on the ceiling of the Sistine Chapel, one might give this gift to a robot.

Finally, I want to put a link to a somewhat more scientific treatment of my ideas on consciousness, submitted to the *Journal of Cognitive Psychology* from my last three posts: the link.

This post has been translated into Russian by Andy Michael: CLICK HERE. Thank you Andy.