Is Neuroscience a Bigger Threat than Artificial Intelligence?

IBM's *Jeopardy* winning computer Watson is a serious threat, not just to the livelihood of medical diagnosticians, but to other professionals who may find themselves going the way of welders. Besides its economic threat, the advance of AI seems to pose a cultural threat: if physical systems can do what we do without thought to give meaning to their achievements, the conscious human mind will be displaced from its unique role in the universe as a creative, responsible, rational agent.

But this worry has a more powerful basis in the Nobel Prize winning discoveries of a quartet of neuroscientists—Eric Kandel, John O'Keefe, Edvard, and May-Britt Moser. For between them they have shown that the human brain doesn't work the way conscious experience suggests at all. Instead it operates to deliver human achievements in the way IBM's Watson does. Thoughts with meaning have no more role in the human brain than in artificial intelligence.

Consciousness tells us that we employ a *theory of mind*, both to decide on our own actions and to predict and explain the behavior of others. According to this theory there have to be particular belief/desire pairings somewhere in our brains working together to bring about movements of the body, including speech and writing. Which beliefs and desires in particular? Roughly speaking it's the contents of beliefs and desires—what they are about—that pair them up to drive our actions. The desires represent the ends, the beliefs record the available means to attain them. It is thus that we give meaning to our actions, and make sense of what others do.

Cognitive scientists have extracted the theory of mind from conscious introspection and elaborated it in a flow chart, a "boxology" of how the mind works. There's what cognitive scientists call a 'desire box' and a 'belief box' inside our heads, one "containing" statements describing the goal or aims, and the other statements about facts relevant to their attainment, information *about* means.

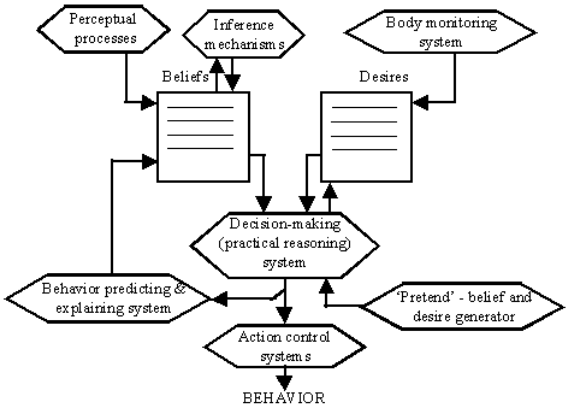

Figure 1

From Shawn Nichols, with permission.

Except for a brief period when psychologists embraced a wrongheaded behaviorism, the theory of mind everyone shares drove the 20th century research programs of child psychology and psychiatry, cognitive science, evolutionary anthropology, and neuroscience.

Several sources of evidence suggest that we have an innate mind-reading ability more powerful than other primates. It's an ability to track other people's actions that is triggered soon after birth. Child psychologists have established its operation in pre-linguistic toddlers, while primatologists have shown its absence in other primates even when they exceed infants in other forms of reasoning. Social psychologists have established deficiencies in its deployment among children on the Autism spectrum. fMRI and transcranial magnetic stimulation studies have localized a brain region that delivers this mind-reading ability. Evolutionary anthropology, game theory and experimental economics have established the indispensability of powerful mind reading for the cooperation and collaboration that resulted in *Hominin* genus's rapid ascent of the African savanna's food chain.

Humanity has converted this innate mind reading ability into a theory of mind with a powerful but nearly invisible role in our understanding of human

action. We've built our cultural, legal and political institutions on this theory that people's actions are caused by choices made rational in the light of their beliefs and their desires.

The theory of mind we all carry around with us almost since birth creates our craving for stories with plots, narratives about human achievements, with intriguing beginnings, tension filled middles, satisfying denouements. The taste for narrative driven by the theory of mind fosters our demands for history and for historical fiction—for stories--true or artfully created.

But here's the problem: the theory of mind we call carry around with us and use every day has no basis in what neuroscience—Nobel Prize winning neuroscience--tell us about how the brain works. Neuroscience has revealed that the theory is quite as much of a dead end as Ptolemaic astronomy. It's been around for such a longtime only because it was the predictive device natural selection came up with, in spite of being fundamentally mistaken about how things were really arranged.

Eric Kandel, John O'Keefe, May-Britt and Edvard Moser, won Nobel prizes in 2000 and 2014 for experiments that showed exactly how the brain records information. Their work revealed it doesn't do it in anything like the way the theory of mind says it does—in statements that represent the way the world is (beliefs) and ones that represent the way we want things to be (desires). This research program began with HM, the patient famous for being unable to acquire or store beliefs because of a lobotomy that went wrong and destroyed his hippocampus. The irony of this research is that it ended up showing that no ones' brain acquires, stores, and uses information in the form of beliefs and desires.

It was Kandel who located and identified the common electrochemical mechanism that all neural circuitry employs to learn and retain information. Then O'Keefe and the Mosers showed how brains record and store the information we mistakenly describe as beliefs about the world in which we find ourselves.

Because the experiments are invasive, they had to use rats. Because rats can't talk they had to identify statements unambiguously attributable to rats as their beliefs. Statements about location, direction, speed, smells, rewards fill the bill.

O'Keefe and the Mosers found the exact neuronal circuitry's firing patterns in the entorhinal cortex and the hippocampus that record and transmit information about the rat's location in space, direction and speed of motion. Subsequent work has shown how neuronal circuitry in the hippocampus moves this information to long-term storage in the neocortex--by "sharp wave ripples," electrochemical pulses that retain the original pattern of electrochemical excitation in the hippocampus but compress it a hundred fold.

O'Keefe and the Mosers learned how they could read off the rat's location, direction, speed and its environment from the firing of particular neuronal circuits in the entorhinal cortex. The Mosers correlated specific locations of the rat and landmarks in its cage with specific neuronal circuits distributed around the entorhinal cortex. Then they could *interpret* the firings as a correct representation, a map for them, of where the rat is, where it's going and what's in the cage. They could read off the rat's location without watching the rat at all! But note, neither the rat nor any part of its brain constructs a map from the neural firings. It's not giving the neural circuits *content*, treating them as containing statements *about* where the rat is. Experimenters decode firing patterns. Rats don't. They 're just driven by them. Firings are all the same, all over the brain—rat and human. What makes some neural firings into location-recorders and other firings into odor-recorders is just their place in the causal chain, the pathway to further behavior. Rats choose among alternative pathways as a result of neural firings produced by previous experience. But it's not because these neuron circuits contain statements about anything. The neurons don't *represent* to the rat the way it's world is arranged. So they don't work any thing like the way beliefs have to work, pairing up with desires via shred content about means and ends. That goes for our neuronal circuits, assemblies, modules, region, too.

Of course you could argue that what Nobel Prize winning research shows about rats is irrelevant to humans. But you'd be flying in the face of clinical evidence about human deficits and disorders, anatomical and physiological identities between the structure of rat and human brains, and the detailed molecular biology of learning and information transmission in the neuronal circuitry of both us and

Rattus rattus, the very reasons neuroscientists interested in human brains have invested so much time and effort in learning how rat brains work. And won Nobel Prizes for doing it.

But conscious experience is continually shouting out that belief/desire psychology is exactly how the mind does work. Introspection all by itself seems to refute the notion that we don't have beliefs and desires with content that represent what we want and what's available to attain our wants.

Alas, ever since Freud psychologists have diagnosed the illusions, delusions and confabulations in the mind's eye and the mind's ear, in the flow of experiences, feelings and sensations passing through consciousness. The theory of mind is just another one of these illusions, useful for survival and success in the Pleistocene, but a blunt instrument of limited predictive and explanatory power. It emerged out of the more fundamental mind-reading ability we share with other species and used to track predators and prey, threats and opportunities. That undoubtedly inborn ability combined with our unique gift, language, to generate the theory of mind. By colonizing consciousness spoken language turned it into a monologue of silent speech, tricking us that the meaning of spoken words is given by thoughts' *content* when its just silent sounds passing through consciousness. Neuroscience shows that that in our brains the neural circuits neither have nor need *content* to do their jobs.

What does all this *mean*? Watson may beat us at *Jeopardy*, but we are convinced we have something AI will always lack: We are agents in the world, whose decisions, choices, actions are made meaningful by the content of the belief/desire pairings that bring them about. But what if the theory of mind that underwrites our distinctiveness is build on sand, is just another useful illusion foisted upon us by the Darwinian processes that got us here? Then it will turn out that neuroscience is a far greater threat to human distinctiveness than AI will ever be.

Alex Rosenberg