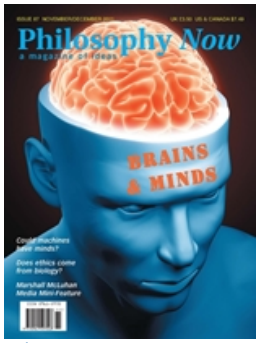
Nov/Dec 2011



Enlarge cover

Back Issues

Podcasts

Search

Forum

Events

Links

Books

Free Articles

Webfeed

FOLLOW US ON

Most Read | Most Discussed | Most Emailed

1. Philosophy of Mind: An Overview

2. Against Stupidity

3. The Death of Postmodernism And Beyond

4. Hawking contra Philosophy

5. News: November/December 2011

Brains & Minds

The Minds of Machines

Print

Email

Discuss

Share

Namit Arora considers the complexity of consciousness and its implications for artificial intelligence.

As a graduate student of computer engineering in the early 90s, I recall impassioned late night debates on whether machines can ever be intelligent – meaning, possessing the cognition, common sense, and problem-solving skills of ordinary humans. Scientists and bearded philosophers spoke of 'humanoid robots'. Neural network research was hot, and one of my professors was a star in the field. A breakthrough seemed inevitable and imminent. Still, I felt certain that Artificial Intelligence (AI) was a doomed enterprise. I argued out of intuition, from a sense of the immersive nature of life: how much we subconsciously acquire and call upon to get through life; how we arrive at meaning and significance not in isolation but through embodied living; and how contextual, fluid, and intertwined these things are with our moods, desires, experiences, selective memory, physical body, and so on. How can we program all this into a machine and have it pass the Turing test, so that we couldn't distinguish its responses from those of a human? How could a machine that did not care about its own existence ever behave as humans do? In hindsight, it seems fitting that I was then also drawn to Dostoevsky, Camus and Kierkegaard.

My interlocutors countered that although extremely complex, the human brain is clearly an instance of matter amenable to the laws of physics. They posited a reductionist and computational approach to the brain that many, including Steven Pinker and Daniel Dennett, continue to champion today. (Recently Dennett declared, "We are robots made of robots made of robots made of robots." – see 'Daniel Dennett Explains How People Are Like Robots', *Bigthink.com*, 9 Mar 2009.) Our intelligence, and everything else that informs our being in the world, had to be somehow coded into our brain's circuitry – including the great many symbols, rules, and associations we rely on to get through a typical day. Was there any reason why we couldn't decode this, and reproduce intelligence in a machine some day? Couldn't a future supercomputer mimic our entire neural circuitry and be as smart as us?

Today's supercomputers are ten million times faster than those of the early 90s. But despite the big advances in computing, AI has fallen woefully short of its ambition and hype. Instead, we have 'expert' systems that process predetermined inputs in specific domains, perform pattern matching and database lookups, and algorithmically learn to adapt their outputs. Examples include chess software, search engines, speech recognition, industrial and service robots, and traffic and weather forecasting systems. Machines have done well with tasks that we ourselves can pursue algorithmically (ie, in a series of small specifiable steps) – as in searching for the word 'ersatz' in an essay, making cappuccino, or restacking books on a library shelf. But so much else that defines our intelligence remains well beyond machines – such as using our creativity and imagination to understand *new* contexts and their significance, or figuring out how and why*new* sensory stimuli are relevant or not. Why is AI in such a brain-dead state? Is there any hope for it? Let's take a closer look.

The Death of Symbolic AI

René Descartes held that science and math would one day explain everything in nature. Early AI researchers embraced Hobbes' view that reasoning was calculating, Leibniz's idea that all knowledge could be expressed as a set of primitives [basic ideas], and Kant's belief that all concepts were rules. At the heart of Western rationalist metaphysics – which shares a remarkable continuity with ancient Greek and Christian metaphysics – lay Cartesian mind-body dualism. This became the dominant inspiration for early AI research. Early researchers pursued what is now known as 'symbolic AI'. They assumed that our brain stored discrete thoughts, ideas, and memories at discrete points, and that information is 'found' rather than 'evoked' by humans. In other words, the brain was a repository of symbols and rules which mapped the external world into neural circuits. And so the problem of creating AI was thought to boil down to creating a gigantic knowledge base with efficient indexing, ie, a search engine extraordinaire. That is, the researchers thought that a machine could be made as smart as a human by storing context-free facts, and rules which would reduce the search time effectively. Marvin Minsky of MIT's AI lab went as far as claiming that our common sense could be produced in machines by encoding ten million facts about objects and their functions.

It is one thing to feed millions of facts and rules into a computer, another to get it to recognize their significance and relevance. The 'frame problem', as this last problem is called, eventually became insurmountable for the 'symbolic AI' research paradigm. One critic, Hubert L. Dreyfus, expressed the problem thus: "If the computer is running a representation of the current state of the world and something in

the world changes, how does the program determine which of its represented facts can be assumed to have stayed the same, and which might have to be updated?" ('Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian').

GOFAI – Good Old Fashioned Artificial Intelligence – as symbolic AI came to be called, soon turned into what philosophers of science call a degenerative research program – reduced to reacting to new discoveries rather than making them. It is unsettling to think how many prominent scientists and philosophers held (and continue to hold), such naïve assumptions about how human minds operate. A few tried to understand what went wrong and looked for a new paradigm for AI. No longer could they ignore the withering critiques of their work by philosophers such as Dreyfus, who drew inspiration from the radical ideas of the German philosopher Martin Heidegger (1889-1976). It began dawning on them that humans were far more complex than they had earlier allowed for, with our subconscious familiarity and skillful coping with the world, nonlinear decision-making, ability to assess and adapt to new situations; and the role of things like purpose, intention, and creativity, that shaped, and were shaped by their organization of the world.

Humanity According To Heidegger

A hammer, Heidegger pointed out, cannot be represented by just its physical features and function, detached from its relationship to nails and the anvil, the experience and skill in hammering of the person using it, or the hammer's role in building fine furniture and comfortable houses. Merely associating facts, values or functions with objects cannot capture the human idea of an object, with its particular role in the meaningful organization of the world as we experience it. As Professor William Blattner writes in *Heidegger's Being and Time* (2006), "Heidegger argues that meaningful human activity, language, and the artifacts and paraphernalia of our world not only make sense in terms of their concrete social and cultural contexts, but also are what they are in terms of that context." (pps.4-5).

Consider hi-fi speakers. One way to represent them, in the manner of rationalists, is as objects with physical properties – shape, dimensions, color, material, attached wires – to which are then assigned a value or function. But this is not how we actually experience music speakers. We experience them as inseparable from the act of listening to music, from the ambience they add to our living room, from their impact on our mood, and so on. We do not understand these objects as context-free, object-value pairs: we understand them through our context-laden use of them. When someone asks us to describe our speakers, we have to pause and think about their physical attributes.

According to Heidegger, writes Professor Blattner:

> "The philosophical tradition has misunderstood human experience by imposing a subject-object schema upon it. The individual human being has traditionally been understood as a rational animal, that is, an animal with cognitive powers, in particular the power to represent the world around it … the notion that human beings are persons and that persons are centers of subjective experience has been broadly accepted … Where the tradition has gone wrong is that it has interpreted subjectivity in a specific way, by means of concepts of 'inner' and 'outer,' 'representation' and 'object'… [which] dominates modern philosophy, from Descartes through Kant through Husserl." (*Heidegger's Being and Time*, p.9)

So in many ways Heidegger stood opposed to the entire edifice of Western philosophy. According to him, the Western philosophical tradition, "has been focused on self-consciousness and moral accountability, in which we experience ourselves as distinct from the world and others." Such 'subject-object dualism' dominates modern science, but fails to describe how humans relate to the world in their experience of it, which is quite holistic. Heidegger claimed that the subject-object model of experience, in which we see ourselves as distinct from the world and others, "does not do justice to our experience, that it forces us to describe our experience in awkward ways, and places the emphasis in our philosophical inquiries on abstract concerns and considerations remote from our everyday lives." (p.48.) As Heidegger contends, "we are disclosed to ourselves more fundamentally than in cognitive self-awareness or moral accountability… Our being is an issue for us, an issue we are constantly addressing by living forward into a life that matters to us." For Heidegger, our being in the world is "more basic than thinking and solving problems; it is not representational at all." For instance, when we are absorbed in work, using familiar pieces of equipment, "the distinction between us and our equipment – between inner and outer – vanishes." Or as Prof Blattner says, Heidegger "argues that our fundamental experience of the world is one of familiarity. We do not normally experience ourselves as subjects standing over against an object, but rather as at home in a world we already understand. We act in a world in which we are immersed. We are not just absorbed in the world, but our sense of identity, of who we are, cannot be disentangled from the world around us. We are what matters to us in our living; we are implicated in the world." (p.12)

In other words, it makes no sense to believe that our minds are built on basic, atomic, context-free sets of facts and rules, with objects and predicates, and discrete storage and processing units. This is why the methods of natural science, which look for structural primitives such as particles and forces, fail to describe our experience. Therefore, contrary to the implicit beliefs of much Western philosophy and AI research, a 'computational' theory of the mind may be impossible. Isn't our common sense "a combination of skills, practices, discriminations, etc, which are not intentional states, and so, *a fortiori*, do not have any representational content to be explicated in terms of elements and rules?" as Hubert L. Dreyfus and Stuart E. Dreyfus asked in *Making a Mind vs. Modeling the Brain: AI Back at a Branchpoint*. The older Wittgenstein agreed, adding in 1948 in his *Last Writings on the Philosophy of Psychology*: "Nothing seems more possible to me than that people some day will come to the definite opinion that there is no copy in the… nervous system which corresponds to a particular thought, or a particular idea, or [a particular] memory."

Intelligence Evolves

A conceptual advance for AI came when some researchers recognized that a computer's model of the world was not real. By comparison, the human 'model' of the world was the world itself, not a static description of it. What if a robot too used the world as its model, "continually referring to its sensors rather than to an internal world model"? (Hubert L. Dreyfus, *What Computers Still Can't Do*). However, this approach worked only in micro-environments with a limited set of features which could be recognized by its sensors. The robots did nothing more sophisticated than ants. As in the past, no one knew how to make the robots learn, or respond to a change in context or significance. This was the backdrop against which AI researchers began turning away from symbolic AI to simulated neural networks, with their promise of self-learning and establishing relevance. Slowly but surely, the AI community began embracing Heideggerean insights about consciousness.

Starting with a blank slate (unlike humans), machine neural networks attempt to simulate brains using a connectionist approach capable of continually adapting its structure based on what it processes and learns. As the Dreyfuses say in *Making a Mind vs. Modeling the Brain*, in symbolic AI, a feature "is either present or not. In the [neural] net, however, although certain nodes are more active when a certain feature is present in the domain, the amount of activity varies not just with the presence or absence of this feature, but is affected by the presence or absence of other features as well." Here, learning is guided using one of three paradigms: supervised learning in controlled domains; unsupervised learning or reinforcement learning based on optimizing certain outcomes.

But the results are not promising. Supervised learning, for instance, remains mired in very basic problems – such as the neural net's inability to generalize predictably in terms of categories intended by the trainer (except for toy problems which leave little room for ambiguity). For example, a net trained to recognize palm trees in photos taken on a sunny afternoon may learn to pick them out by generalizing on their shadows, and thus fail to detect any trees in photos from an overcast day. The sample size can be enlarged; but the point is that the trainer doesn't know what the net is precisely training itself to do. Another neural net trained to recognize speech may crash when it encounters a metaphor – say, 'Sally is a block of ice'. Outside its training domain, the net is also unable to recognize other contexts, and therefore cannot know when it is not appropriate to apply what it has learned – problems that humans dynamically solve using their widely-comprehending consciousnesses, involving social skills, biological drives, imagination, and more.

Reinforcement learning has its own pitfalls. For instance, what is an objective measure of reinforcement? Even if we take a simplistic view that humans act so as to maximize satisfaction and assign a 'satisfaction score' to all foreseeable outcomes, we need some way to model and artificially reproduce how 'satisfaction' may be affected by our moods, desires, body aches, etc, as well as modelling their correlation with inputs in a diversity of situations (weather, familiar faces, noise, motion, etc). But does anyone know what model rules, if any, humans obey in their daily behavior? The Dreyfuses sum it up: "If [a simulated neural net] is to learn from its own 'experiences' to make associations that are human-like rather than be taught to make associations which have been specified by its trainer, it must also share our sense of appropriateness of outputs, and this means it must share our needs, desires, and emotions, and have a human-like body with the same physical movements, abilities and possible injuries." (*Making a Mind vs. Modeling the Brain*).

In other words, the success of neural nets will depend not only on our understanding of how we breathe significance and meaning into our world (which was Heidegger's endeavour), and finding a way to capture this understanding in the language of machines: in order to have a shot at behaving like humans, these nets also need to come into a social world similar to that of humans and project themselves in time the way humans do with their physical bodies. How to achieve any of this is not even remotely clear to anyone, nor is it clear that these things are even amenable to modeling on digital computers. To insist otherwise is not only an article of faith, it also seems to me increasingly obtuse and wild.

Namit Arora lives in the San Francisco Bay area and is the creator of Shunya*, an internet photo journal.*