Awakening Self-reflective Intelligence

Paul Hague

March 2017

In 1950, Alan Turing wrote an article published in *Mind*, in which he asked the question, "Can machines think?" He thought that by the end of the century we would be able to say yes to this question "without expecting to be contradicted". To test this prediction he devised an 'imitation game', known today as the 'Turing test', in which observers, asking the same questions to humans and machines, would not be able to distinguish who was which from their responses.

Then in 1955, John McCarthy coined the term *artificial intelligence*, convening a conference at Dartmouth College the following year to explore the belief that human intelligence "can be so precisely described that a machine can be made to simulate it". For myself, I began to ponder this cognitive issue in 1964, when I wrote a program in Fortran on an IBM 7094 to calculate the roots of a quadratic equation.

Now, while no one has yet won the $100,000 prize offered by Hugh Loebner for the first computer to pass the Turing test, computer scientists are all abuzz that this long-awaited breakthrough is about to happen at a technological singularity in time. After that, the world will be a completely different place, leading to a major existential, psychological, and economic crisis.

For what will then be humanity's purpose in life? How will we joyfully earn our livelihood, making a worthwhile contribution to the well-being of our fellow humans in our daily lives? These were questions that I asked myself in 1980, when I was engaged in developing an innovative marketing programme for decision support systems for IBM in London. Not knowing how to answer them—not able to take responsibility for the changes I was helping to bring about in the world—I left my business career to explore them for myself, adopting a liberating work ethic. I was driven by an incredible surge of creative energy pouring through me, which my education as a mathematician had not prepared me to understand.

Today, I do. As I can now see, the convergent, holistic powers of evolution have led me to reverse Turing's imitation game. Starting afresh at the very beginning—at the Origin of the Universe—I embarked on a thought experiment in which I imagined that I was a computer that switched itself off and on again so that it had no programs within it, not even a bootstrap program to load the operating system.

Adapting the method of *reductio ad absurdum* in mathematics, this machine then had the task of solving the ultimate problem in human learning, known as the Theory of Everything, which thinkers have been searching for since Roger Bacon in the 1200s. If a human rather than a machine could integrate all knowledge in all cultures and disciplines at all times into a coherent whole, like the Internet, then I would discover the essential difference between humans and machines, confirming my intuition that we humans have far more potential to awaken Self-reflective Intelligence than machines will ever have.

But what to do now? There is the most gigantic gulf between depth psychologists and DeepMind Technologies, for instance, both attempting to develop a systemic science of the mind, foreseen by William James and Gottfried Wilhelm Leibniz on the two sides of the divide. George Boole made an initial attempt to bridge this gap in 1856 with his *Laws of Thought*, opened up again by Bertrand Russell and Gottlieb Frege, who agreed in 1902 that mathematical logic has nothing to do with psychology.

Healing this deep split in the cultural psyche shows that psychology, rather than physics, is the primary science, on which all humanities and sciences are built. For our minds create our reality and govern our behaviour, guided by the creative power of Life emanating from Reality, the Divine Source of everything.

So if we are we to confront the dangers from artificial general intelligence and the other existential risks that humanity faces today, we need to cocreate a cultural environment in which it is socially acceptable to engage in objective self-inquiry, breaking the scientific and religious taboos on mapping our inner worlds.

This is absolutely essential, for Martin Rees—Astronomer Royal and former President of the Royal Society—writes in *Our Final Century* that while science and technology have provided many of us with the most amazing creature comforts during the last century or two, "The 'downside' from twenty-first century technology could be graver and more intractable than the threat of nuclear devastation." Specifically, he cites the potential dangers of nanotechnology, genetic engineering, and artificial intelligence, saying, "A superintelligent machine could be the last invention that humans need ever make." Similarly, Stephen Hawking, long searching for the Theory of Everything, told the BBC on 2nd December 2014, "The development of full artificial intelligence could spell the end of the human race."

Today, centres at Oxford and Cambridge Universities are addressing these existential risks, a term coined by Nick Bostrom, Director of the Future of Humanity Institute in Oxford, originally funded by James Martin, a fellow alumnus of IBM, whose books much influenced my work as a systems engineer in an IBM sales office in the early 1970s and my later researches. And at Cambridge, Martin Rees has cofounded the Centre for the Study of Existential Risk (CSER), as an interdisciplinary research centre focused on the study of human extinction-level risks that may emerge from technological advances.

However, from what I have read, these institutions are not yet going nearly far enough. For example, in *Superintelligence*, Nick Bostrom gives several examples of 'superhuman' abilities in computers, such as the ability to beat human experts in games like Chess, Othello, and Jeopardy! Since he wrote this book, DeepMind's AlphaGo has defeated a 9-dan Go champion using a deep learning technique. These games are well-structured, albeit of immense complexity.

Yet, in *Decision Support Systems* in 1978, Michael S. Scott Morton and Peter G. W. Keen categorized human tasks as structured, semi-structured, and unstructured, the last requiring intelligence and intuition; they cannot be automated. It was this classification that led me to study how humans could perform unstructured tasks in 1980 and why they cannot be automated. What I have discovered is that while human learning is an accumulative, evolutionary, pattern-making process, deep learning requires us to forget what we have learnt, essentially because existing cognitive structures can inhibit further growth, preventing us from awakening Self-reflective Intelligence, necessary to solve today's unprecedented global crisis.

So while DeepMind's motto is 'Solve intelligence. Use it to make the world a better place,' its recent paper 'Overcoming catastrophic forgetting in neural networks'—showing how algorithmic computers can optimize the performance of playing several Atari games at once by mimicking synaptic consolidation in the brain—doesn't fully help us to understand human intelligence vis-à-vis 'machine intelligence'.

Indeed, there is no area of human endeavour that is more beset with confusion and uncertainty and what this might mean for our children and grandchildren's future. Inspired by DeepMind's recent paper, co-authored by James Kirkpatrick, Ian Sample, *The Guardian*'s science editor, wrote an opinion piece on 15th March titled 'AI is getting brainier: when will the machines leave us in the dust?' Three days later, Greg Jericho wrote another article in *The Guardian* with the rubric 'An automated world is coming and managing the unemployment fallout won't be easy', adding the subtitle 'If automation pushes joblessness to 20%, what happens to those who are left behind?'

This, for me, is the central issue of our times, as it has been since 1979, when I was unable to answer customer executives' questions on the changes we were introducing into the workplace, when giving keynote presentations at IBM's European Education Centre in Belgium. After 37 years of self-inquiry, I have now found satisfactory answers to these existential questions, which I would be delighted to share.