Examining the Work and Its Later Impact

Daniel Dennett is inspired by —

TURING'S "STRANGE INVERSION OF REASONING"

Some of the greatest, most revolutionary advances in science have been given their initial expression in attractively modest terms, with no fanfare. Charles Darwin managed to compress his entire theory into a single summary paragraph that a layperson can readily follow, in all its details:

> If during the long course of ages and under varying conditions of life, organic beings vary at all in the several parts of their organization, and I think this cannot be disputed; if there be, owing to the high geometric powers of increase of each species, at some age, season, or year, a severe struggle for life, and this certainly cannot be disputed; then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite diversity in structure, constitution, and habits, to be advantageous to them, I think it would be a most extraordinary fact if no variation ever had occurred useful to each being's own welfare, in the same way as so many variations have occurred useful to man. But if variations useful to any organic being do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterized. This principle of preservation, I have called, for the sake of brevity, Natural Selection. (*Origin of Species*, end of chapter 4)

Francis Crick and James Watson closed their epoch-making paper on the structure of DNA with the deliciously diffident sentence:

> It has not escaped our notice that the specific pairings we have postulated immediately suggests a possible copying mechanism for the replicating unit of life. (Watson and Crick (1953), p.738)

And Alan Turing created a new world of science and technology, setting the stage for solving one of the most baffling puzzles remaining to science, the mind-body problem, with an even shorter declarative sentence in the middle of his 1936 paper on computable numbers:

> It is possible to invent a single machine which can be used to compute any computable sequence. (Turing (1936), p.241)

Turing didn't just intuit that this remarkable feat was possible; he showed exactly how to make such a machine. With that demonstration the computer age was born. It is important to remember that there were entities called computers before Turing came up with his idea – but they were people, clerical workers with enough mathematical skill, patience, and pride in their work to generate reliable results of hours and hours of computation, day in and day out. Many of them were women.

Dryden Flight Research Center E49-0053 Photographed 10/49
Early "computers" at work. NASA photo

Thousands of them were employed in engineering and commerce, and in the armed forces and elsewhere, calculating tables for use in navigation, gunnery and other such technical endeavours. A good way of understanding Turing's revolutionary idea about computation is to put it in juxtaposition with Darwin's about evolution. The pre-Darwinian world was held together not by science but by tradition: all things in the universe, from the most exalted ('man') to the most humble (the ant, the pebble, and the raindrop), were the creations of a still more exalted thing, God, an omnipotent and omniscient intelligent creator – who bore a striking resemblance to the second-most exalted thing. Call this the trickle-down theory of creation. Darwin replaced it with the bubble-up theory of creation. One of Darwin's nineteenth century critics put it vividly:

> In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that, IN ORDER TO MAKE A PERFECT AND BEAUTIFUL MACHINE, IT IS NOT REQUISITE TO KNOW HOW TO MAKE IT. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the Theory, and to express in a few words all Mr. Darwin's meaning; who, by a strange inversion of reasoning, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all the achievements of creative skill. (MacKenzie, 1868)

It was, indeed, a strange inversion of reasoning. To this day many people cannot get their heads around the unsettling idea that a purposeless, mindless process can crank away through the eons, generating ever more subtle, efficient and complex organisms without having the slightest whiff of understanding of what it is doing.

Turing's idea was a similar – in fact remarkably similar – strange inversion of reasoning. The Pre-Turing world was one in which computers were people, who had to understand mathematics in order to do their jobs. Turing realised that this was just not necessary: you could take the tasks they performed and squeeze out the last tiny smidgens of understanding, leaving nothing but brute, mechanical actions. IN ORDER TO BE A PERFECT AND BEAUTIFUL COMPUTING MACHINE IT IS NOT REQUISITE TO KNOW WHAT ARITHMETIC IS.

What Darwin and Turing had both discovered, in their different ways, was the existence of *competence without comprehension* (Dennett, 2009, from which material in the preceding paragraphs has been drawn, with revisions). This inverted the deeply plausible assumption that comprehension is in fact the *source* of all advanced competence. Why, after all, do we insist on sending our children to school, and why do we frown on the old-fashioned methods of rote learning? We expect our children's growing competence to *flow from* their growing comprehension; the motto of modern education might be: 'comprehend *in order to be* competent' And for us members of *H. sapiens*, this is almost always the right way to look at, and strive for, competence. I suspect that this much-loved principle of education is one of the primary motivators of skepticism about both evolution and its cousin in Turing's world, Artificial Intelligence. The very idea that mindless mechanicity can generate human-level – or divine level! – competence strikes many as philistine, repugnant, an insult to our minds and the mind of God.

Consider how Turing went about his proof. He took human computers as his model. There they sat at their desks, doing one simple and highly reliable step after another, checking their work, writing down the intermediate results instead of relying on their memories, consulting their recipes as often as they needed, turning what at first might appear a daunting task into a routine they could almost do in their sleep. Turing systematically broke down the simple steps into even simpler steps, removing all vestiges of discernment or comprehension. Did a human computer have difficulty telling the number 99999999999 from the number 9999999999? Then, break down the perceptual problem of *recognizing the number* into simpler problems, distributing easier, stupider acts of discrimination over multiple steps. He thus prepared an inventory of basic building blocks from which to construct the universal algorithm that could execute any other algorithm. He showed how that algorithm would enable a (human) computer to compute any function, and noted that:

> The behaviour of the computer at any moment is determined by the symbols which he is observing and his "state of mind" at that moment. We may suppose that there is a bound *B* to the number of symbols or squares which the computer can observe at one moment. If he wishes to observe more, he must use successive observations. .... The operation actually performed is determined .... by the state of mind of the computer and the observed symbols. In particular, they determine the state of mind of the computer after the operation is carried out.

He then noted, calmly:

> We may now construct a machine to do the work of this computer. (p.251)

Right there we see the reduction of *all possible computation* to a mindless process. We can start with the simple building blocks Turing had isolated, and construct layer upon layer of more sophisticated computation, restoring, gradually, the intelligence Turing had so deftly laundered out of the practices of human computers.

But what about the genius of Turing, and of later, lesser programmers, whose own intelligent comprehension was manifestly the source of the designs that can knit Turing's mindless building blocks into useful competences? Doesn't this dependence just re-introduce the trickle-down perspective on intelligence, with Turing in the God role? No less a thinker than Roger Penrose has expressed skepticism about the possibility that Artificial Intelligence could be the fruit of nothing but mindless algorithmic processes.

> I am a strong believer in the power of natural selection. But I do not see how natural selection, in itself, can evolve algorithms which could have the kind of conscious judgements of the validity of other algorithms that we seem to have. (1989, p.414)

He goes on that admit:

> To my way of thinking there is still something mysterious about evolution, with its apparent'groping' towards some future purpose. Things at least *seem* to organize themselves somewhat better than they'ought' to, just on the basis of blind-chance evolution and natural selection. (1989, p.416)

Indeed, a single cascade of natural selection events, occurring over even billions of years, would seem unlikely to be able to create a string of zeroes and ones that, once read by a digital computer, would be an 'algorithm' for 'conscious judgments.' But as Turing fully realised, there was nothing to prevent the process of evolution from copying itself on many scales, of mounting discernment and judgment. The recursive step that got the ball rolling – designing a computer that could mimic any other computer–could itself be reiterated, permitting specific computers to enhance their own powers by *redesigning themselves*, leaving their original designer far behind. Already in 'Computing Machinery and Intelligence,' his classic paper in *Mind*, 1950, he recognised that there was no contradiction in the concept of a (non-human) computer that could learn.

> The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States. (See Suber (2001), unpublished, for a valuable discussion of this passage and the so-called paradox of self-amendment.)

He saw clearly that all the versatility and self-modifiability of human though – learning and re-evaluation and, language and problem-solving, for instance – could in principle be constructed out of these building blocks. Call this the bubble-up theory of mind, and contrast it with the various trickle-down theories of mind, by thinkers from Rene´ Descartes to John Searle (and including, notoriously, Kurt G¨odel, whose proof was the inspiration for Turing's work) that start with human consciousness at its most reflective, and then are unable to unite such magical powers with the mere mechanisms of human bodies and brains.

Turing, like Darwin, broke down the mystery of intelligence (or Intelligent Design) into what we might call atomic steps of dumb happenstance, which, when accumulated by the millions, added up to a sort of pseudo-intelligence. The Central Processing Unit of a computer doesn't *really* know what arithmetic is, or understand what addition is, but it 'understands' the 'command' to add two numbers and put their sum in a register – in the minimal sense that it reliably adds when thus called upon to add and puts the sum in the right place. Let's say it *sorta* understands addition. A few levels higher, the operating system doesn't *really* understand that it is checking for errors of transmission and fixing them but it *sorta* understands this, and reliably does this work when called upon. A few further levels higher, when the building blocks are stacked up by the billions and trillions, the chess-playing program does't *really* understand that its queen is in jeopardy, but it *sorta* understands this, and IBM's Watson on Jeopardy *sorta* understands the questions it answers.

Why indulge in this *'sorta'* talk? Because when we analyze – or synthesise – this stack of ever more competent levels, we need to keep track of two facts about each level: what it *is* and what it *does*. What it *is* can be described in terms of the structural organization of the parts from which it is made – so long as we can assume that the parts function as they are supposed to function. What it *does* is some (cognitive) function that it (sorta) performs – well enough so that at the next level up, we can make the assumption that we have in our inventory a smarter building block that performs just that function – sorta, good enough to use. This is the key to breaking the back of the mind-bogglingly complex question of how a mind could ever be composed of material mechanisms.

What we might call the *sorta* operator is, in cognitive science, the parallel of Darwin's gradualism in evolutionary processes. Before there were bacteria there were *sorta* bacteria, before there were mammals there were *sorta* mammals and before there were dogs there were *sorta* dogs, and so forth. We need Darwin's gradualism to explain the huge difference between an ape and an apple, and we need Turing's gradualism to explain the huge difference between a humanoid robot and a hand calculator. The ape and the apple are made of the same basic ingredients, differently structured and exploited in a many-level cascade of different functional competences. There is no principled divid-ing line between a *sorta* ape and an ape. The humanoid robot and the hand calculator are both made of the same basic, unthinking, unfeeling Turing-bricks, but as we compose them into larger, more competent structures, which then become the elements of still more competent structures at higher levels, we eventually arrive at parts so (*sorta*) intelligent that they can be assembled into compe-tences that deserve to be called comprehending. We use the intentional stance (Dennett, 1971, 1987) to keep track of the beliefs and desires (or 'beliefs' and 'desires' or sorta beliefs and sorta desires) of the (sorta-)rational agents at every level from the simplest bacterium through all the discriminat-ing, signaling, comparing, remembering circuits that compose the brains of animals from starfish to astronomers. There is no principled line above which true comprehension is to be found – even in our own case. The small child *sorta* understands her own sentence 'Daddy is a doctor', and I *sorta* understand '$E = mc^{2}$'Some philosophers resist this anti-essentialism: either you believe that snow is white or you don't; either you are conscious or you aren't; nothing counts as an approximation of any mental phenomenon – it's all or nothing. And to such thinkers, the powers of minds are insolu-ble mysteries because they are 'perfect,' and perfectly unlike anything to be found in mere material mechanisms.

We still haven't arrived at 'real' understanding in robots, but we are getting closer. That, at least, is the conviction of those of us inspired by Turing's insight. The trickle-down theorists are sure in their bones that no amount of further building will ever get us to the real thing. They think that a Cartesian *res cogitans*, a thinking thing, cannot be constructed out of Turing's building blocks. And creationists are similarly sure in their bones that no amount of Darwinian shuffling and copying and selecting could ever arrive at (real) living things. They are wrong, but one can appreciate the discomfort that motivates their conviction.

Turing's strange inversion of reason, like Darwin's, goes against the grain of millennia of earlier thought. If the history of resistance to Darwinian thinking is a good measure, we can expect that long into the future, long after every triumph of human thought has been matched or surpassed by 'mere machines', there will still be thinkers who insist that the human mind works in mysterious ways that no science can comprehend.

References

Darwin, C., 1859. On the Origin of Species, John Murray, London.

Dennett, D.C., 1971. Intentional systems. J. Phil., 68, 87–106.

Dennett, D.C., 1987. The Intentional Stance, MIT Press, Cambridge, MA.

Dennett, D.C., 2009. Darwin's Strange inversion of reasoning. PNAS, 106, 10061–10065.

MacKenzie, R.B., 1868. The Darwinian Theory of the Transmutation of Species Examined, Nisbet & Co, London.

Penrose, R., 1989. The Emperor's New Mind, Oxford University Press, Oxford.

Suber, P., (unpublished). Saving Machines From Themselves: The Ethics of Deep Self-Modification, preprint, 30 November 2001. http://www.earlham.edu/~peters/writing/selfmod.htm

Turing, A., 1936. On computable numbers, with an application to the Entscheidungsproblem. Proc. Lond. Math. Soc. 42, 23–265, and erratum (1937) 43, 544–546.

Turing, A., 1950. Computing machinery and intelligence. Mind, LIX, 2236, 433–460.

Watson, J.D. and Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.