

3. Welcome to a world of exponential change

Nick Bostrom

For most of human history, the pace of technological development was so slow that a person might be born, live out a full human life and die without having perceived any appreciable change. In those times, worldly affairs appeared to have a cyclical nature. Tribes flourished and languished, bad rulers came and went, empires expanded and fell apart in seemingly never-ending loops of creation and destruction. To the extent that there was a direction or destination to all this striving, it was commonly thought to lie outside time altogether, in the realm of myth or supernatural intervention.

A present day observer, by contrast, expects to see significant technological change within a time span as short as a decade and much less in certain sectors. Yet although the external factors of the human condition have been profoundly transformed and continue to undergo rapid change, the internal factors – our basic biological capacities – have remained more or less constant throughout history. We still eat, sleep, defecate, fornicate, see, hear, feel, think and age in pretty much the same ways as the contemporaries of Sophocles did. But we may now be approaching a time when this will no longer be so.

The prospect of artificial intelligence

The annals of artificial intelligence are littered with broken promises. Half a century after the first electric computer, we still have nothing

that even resembles an intelligent machine, if by 'intelligent' we mean possessing the kind of general-purpose smartness that we humans pride ourselves on. Maybe we will never manage to build real artificial intelligence. The problem could be too difficult for human brains ever to solve. Those who find the prospect of machines surpassing us in general intellectual abilities threatening may even hope that is the case.

However, neither the fact that machine intelligence would be scary nor the fact that some past predictions were wrong is good ground for concluding that artificial intelligence will never be created. Indeed, to assume that artificial intelligence is impossible or will take thousands of years to develop seems at least as unwarranted as to make the opposite assumption. At a minimum, we must acknowledge that any scenario about what the world will be like in 2050 that postulates the absence of human-level artificial intelligence is making a big assumption that could well turn out to be false.

It is therefore important to consider the alternative possibility: that intelligent machines will be built within 50 years. We can get a grasp of this issue by considering the three things that are needed for effective artificial intelligence. These are: hardware, software and input/output mechanisms.

The requisite input/output technology already exists. We have video cameras, speakers, robotic arms etc that provide a rich variety of ways for a computer to interact with its environment. So this part is trivial.

The hardware problem is more challenging. Speed rather than memory seems to be the limiting factor. We can make a guess at the computer hardware that will be needed by estimating the processing power of a human brain. We get somewhat different figures depending on what method we use and what degree of optimisation we assume, but typical estimates range from 100 million MIPS to 100 billion MIPS (1 MIPS = one million instructions per second). A high-range PC today has a few thousand MIPS. The most powerful supercomputer to date performs at 260 million MIPS. This means that we will soon be within striking distance of meeting the hardware

requirements for human-level artificial intelligence. In retrospect, it is easy to see why the early artificial intelligence efforts in the 1960s and 1970s could not possibly have succeeded – the hardware available then was pitifully inadequate. It is no wonder that human-level intelligence was not attained using a less-than-cockroach level of processing power.

Turning our gaze forward, we can predict with a high degree of confidence that hardware matching that of the human brain will be available in the foreseeable future. We can extrapolate using Moore's Law, which describes the historical growth rate of computer speed. (Strictly speaking, Moore's Law as originally formulated was about the density of transistors on a computer chip, but this has been closely correlated with processing power.) For the past half century, computing power has doubled every 18 months to two years. Moore's Law is really not a law at all, but merely an observed regularity. In principle, it could stop holding true at any point in time.

Nevertheless, the trend it depicts has been going strong for an extended period of time and it has survived several transitions in the underlying technology (from relays to vacuum tubes, to transistors, to integrated circuits, to very large integrated circuits). Chip manufacturers rely on it when they plan their forthcoming product lines. It is therefore reasonable to suppose that it may continue to hold for some time. Using a conservative doubling time of two years, Moore's law predicts that the upper-end estimate of the human brain's processing power will be reached before 2020. Since this represents the performance of the best supercomputer in the world, one may add a few years to account for the delay that may occur before that level of computing power becomes available for doing experimental work in artificial intelligence. The exact numbers don't matter much here. The point is that human-level computing power probably has not been reached yet, but almost certainly will be attained well before 2050.

This leaves the software problem. It is harder to analyse in a rigorous way how long it will take to solve that problem. (Of course, this holds equally for those who feel confident that artificial

intelligence will remain unobtainable for an extremely long time – in the absence of evidence, we should not rule out either alternative.) Here we will approach the issue by outlining just one approach to creating the software, and presenting some general plausibility arguments for how it could work.

We know that the software problem can be solved in principle. After all, humans have achieved human-level intelligence, so it is evidently possible. One way to build the requisite software is to figure out how the human brain works, and copy nature's solution.

It is only relatively recently that we have begun to understand the computational mechanisms of biological brains. Computational neuroscience is about 15 years old as an active research discipline. In this short time, substantial progress has been made. We are beginning to understand early sensory processing. There are reasonably good computational models of the primary visual cortex, and we are working our way up to the higher stages of visual cognition. We are uncovering what the basic learning algorithms are that govern how the strengths of synapses are modified by experience. The general architecture of our neuronal networks is being mapped out as we learn more about the interconnectivity between neurones and how different cortical areas project onto one another. While we are still far from understanding higher-level thinking, we are beginning to figure out how the individual components work and how they are hooked up.

Assuming continuing rapid progress in neuroscience, we can envision learning enough about the lower-level processes and the overall architecture to begin to implement the same paradigms in computer simulations. Today, such simulations are limited to relatively small assemblies of neurones. There is a silicon retina and a silicon cochlea that do the same things as their biological counterparts. IBM's 'Blue Brain Project' aims to create an accurate software replica of a neocortical column by 2008. Simulating a whole brain will of course require enormous computing power; but as we saw, that capacity will be available within a couple of decades.

The product of this biology-inspired method will not be an explicitly coded mature artificial intelligence. (That is what the so-

called classical school of artificial intelligence tried unsuccessfully to do.) Rather, it will be a system that has the same ability as a toddler to learn from experience and to be educated. The system will need to be taught in order to attain the abilities of adult humans. But there is no reason why the computational algorithms that our biological brains use would not work equally well when implemented in silicon hardware.

The promise of nanotechnology

In 2005, Europe, the US and Japan spent approximately one billion US dollars each on nanotechnology in public funding, a tenfold increase since 1997.¹ ‘Nanotechnology’ has become a buzzword. Putting it into a grant application can greatly increase its chance of being funded – as Oxford Professor George Smith jokes, ‘nano is from the Greek verb meaning “to attract research funding”’.

The word was coined by Dr Erik Drexler and popularised in his 1986 book *Engines of Creation*.² Drexler published detailed technical analyses arguing for the feasibility of building molecular machines to atomic precision.³ Such machine-phase nanotechnology, in its mature form, will give humanity unprecedented control of the structure of matter. In many respects, it will transform manufacturing into a software problem. In Drexler’s vision, nanotech construction devices would build objects one molecule at a time, and billions of such devices working in parallel would be able to construct atomically almost-perfect objects of arbitrary size. Applications would include:

- extremely fast computers
- lighter, stronger materials (a strong enabling factor for space technology)
- clean, efficient manufacturing processes of most products
- cheap solar energy production, and the ability to actively scrape excessive CO₂ out of the atmosphere
- desktop manufacturing devices with near-universal capabilities

- tiny medical robots that could enter individual cells and perform molecular-level repair, eliminating most disease and ageing, and making it possible to upload human minds to computers (this would be a second possible route to human-level artificial intelligence).

Drexler noted that a technology this powerful could also be used with devastating results for destructive ends. He worried especially about dangerous arms races, new weapons of mass destruction that could be used by terrorists and rogue states, and mind-control technologies that could be used by bad governments to oppress their populations. Nevertheless, Drexler argued, attempts to prevent the development of nanotechnology would necessarily fail and would on balance increase the dangers.

Although Drexler helped create the enthusiasm for the field of nanotechnology that has resulted in the recent funding boom, he has subsequently been sidelined by the mainstream community of nanoscientists because his vision runs too far ahead of the experimental work that is currently being done in labs and the applications that are immediately on the horizon. Another reason for Drexler's marginalisation is the fear felt by some nanotechnologists that the future dangers to which he drew attention could fan public opposition to nanotechnology, resulting in a loss of funding. One Nobel laureate, Richard Smalley, declared – without offering any technical argument – that Drexler's vision was physically impossible and went on to accuse Drexler of 'scaring our children',⁴

More recently, there are some signs that the Drexlerian vision of nanotechnology might be poised for a comeback, thanks partly to rapid scientific progress in the field and new computer modelling studies that seem to support the feasibility of molecular machine systems. Policy-makers are already concerned with the need to examine the far-reaching ethical and social implications that nanotechnology will have once it is fully developed. For example, a nanotechnology bill signed into law by President Bush in late 2003 requires that the programme ensure

that ethical, legal, environmental, and other appropriate societal concerns, including the potential use of nanotechnology in enhancing human intelligence and in developing artificial intelligence which exceeds human capacity, are considered during the development of nanotechnology.⁵

Some 3 per cent of the budget of the Human Genome Project was devoted to studying the ethical, legal and social issues (ELSI) around the availability of genetic information. It looks like nanotechnology is set to continue this new trend of including the social science and the humanities in major technological research programmes. Such anticipatory ELSI research is a new phenomenon, and its long-term effects remain to be seen.

Convergence and the singularity

The concept of ‘converging technologies’ stems from a 2002 report sponsored by the US National Science Foundation (NSF), and edited by Mihail Roco and William Bainbridge:

In the early decades of the twenty-first century, concentrated efforts can unify science based on the unity of nature, thereby advancing the combination of nanotechnology, biotechnology, information technology, and new technologies based in cognitive science. With proper attention to ethical issues and societal needs, converging technologies could achieve a tremendous improvement in human abilities, societal outcomes, the nation’s productivity, and the quality of life.⁶

The phrase ‘converging technologies’ refers to the synergistic combination of four major provinces of science and technology, known in short as ‘NBIC’. These are (a) nanoscience and nanotechnology; (b) biotechnology and biomedicine, including genetic engineering; (c) information technology, including advanced computing and communications; and (d) cognitive science, including cognitive neuroscience. The idea is that as these four areas develop they will

join to create a more integrated approach to science and technology, where, for instance, the boundary between biotechnology and nanotechnology dissolves. The NSF report describes how dramatic new capabilities would result and could be used to enhance human capacities.

It has been said that most people overestimate how much technological progress there will be in the short term and underestimate how much there will be in the long term. There is usually a long lag time between proof-of-concept in some laboratory and the time when actual products begin to have a significant impact in the market. Many a seemingly good idea never pans out. Hot technological fields usually yield a lot of hype.

The world economy is doubling every 15 years. Particular technological areas exhibit faster growth. Ray Kurzweil, the American inventor and technology forecaster, has documented many technological areas, including computing, data storage, gene sequencing, brain mapping and others, where progress is currently occurring at a rapid exponential pace. Exponential growth starts slow and then becomes very fast. Here is a classic problem that illustrates this:

The water lilies in a pond double every day. It takes two weeks before the lilies cover the whole pond. How long did it take before they covered half of the pond?

The answer, of course, is that the exponentially growing lily population covered half the pond on day 13, one day before it doubled again to cover the whole pond. Kurzweil argues that we intuitively tend to think of progress as linear while in reality it is exponential, and that many people will be surprised to find how rapidly things develop over the coming decades. Kurzweil believes that we will not experience 100 years of progress in the twenty-first century – it will be more like 20,000 years of progress (at today's rate).⁷ This is because our ability to invent new things is itself improving, through advances in scientific instrumentation, methodology and computing.

The ‘singularity’ is a hypothetical point in the future where the rate of technological progress becomes so rapid that the world is radically transformed virtually overnight. The only plausible scenario in which such a singularity could occur is through the development of machine intelligence. One might imagine that machines will at some point come to significantly surpass biological human beings in general intelligence, and that these machines will be able to apply their intelligence to rapidly improve themselves so that within short order they become superintelligent. Superintelligent machines would then be able to rapidly advance all other fields of science and technology. Among the many other things that would become possible is the uploading of human minds into computers, and dramatic modification or enhancement of the biological capacities of human beings that remain organic.

It is of course an open question whether a singularity will ever occur. It is possible that there will never be a point where progress becomes as rapid as the singularity hypothesis postulates. Even if there were to be a singularity at some point, it is very difficult to predict how long it would take to get there, although some have argued that it is more likely than not that we will have superintelligent machines before the middle of the twenty-first century.⁸

What is a policy-maker to do in light of all these possibilities? A first priority is to abandon the unquestioning assumption that human nature and the human condition will remain fundamentally unchanged throughout the current century. A second is to develop better techniques for long-range planning and horizon-scanning. Such techniques are already used in some policy decisions, for example, in arguments about the importance of reducing global warming. Yet once we consider the bigger picture, we may feel that the risks of global warming are dwarfed by other risks that our technological advances will create over the coming several decades.⁹ Perhaps we ought to spend a fraction of the money and effort currently devoted to the problem of climate change to thinking about these other risks too.

And in addition to risks, there are also immense opportunities. Again, consideration of the big picture can help us spot opportunities for saving lives and improving the quality of life that might otherwise go unnoticed. A massive increase in funding for research to better understand the basic biology of ageing could pay off handsomely if it leads to treatments to intervene in the negative aspects of senescence, allowing men and women to stay healthy and economically productive much longer than is currently possible.

That there will be change is certain, but what the change will be depends in some measure on human choice. In this century we may choose to use our technological ingenuity to unlock our potential in ways that were unimaginable in the past.

Nick Bostrom is Director of the Future of Humanity Institute at the University of Oxford and Chair and co-founder of the World Transhumanist Association. The section of this article on artificial intelligence is based on 'When machines outsmart humans' by Nick Bostrom in Futures 35, no 7 (2003).

Notes

- 1 M Roco and WS Bainbridge (eds), *Nanotechnology: Societal implications maximizing benefits for humanity*. Report of the National Nanotechnology Initiative Workshop (Arlington, VA, 2003).
- 2 KE Drexler, *Engines of Creation: The coming era of nanotechnology* (London: Fourth Estate, 1986).
- 3 KE Drexler, *Nanosystems: Molecular machinery, manufacturing, and computation* (New York: John Wiley & Sons, Inc., 1992).
- 4 E Drexler and R Smalley (2003). 'Nanotechnology: Drexler and Smalley make the case for and against "molecular assemblers"', *Chemical & Engineering News* 81, no 48 (2003).
5. '21st Century Nanotechnology Research and Development Act' (passed 3 Dec 2003) (1, section 2.B.10); see: www.nano.gov/html/news/PresSignsNanoBill.htm (accessed 17 Jan 06).
- 6 MC Roco and WS Bainbridge (eds), *Converging Technologies for Improving Human Performance* (Arlington, VA: National Science Foundation/Department of Commerce-sponsored report, 2002).
- 7 R Kurzweil, 'The law of accelerating returns', KurzweilAI.net, 2001. Available at: www.kurzweilai.net/articles/art0134.html?printable=1 (accessed 7 Jan 2006).

Better Humans?

- 8 See N Bostrom, 'How long before superintelligence?', *International Journal of Futures Studies* 2 (1998), and H Moravec, *Robot: Mere machine to transcendent mind* (New York: Oxford University Press, 1999).
- 9 N Bostrom, 'Existential risks: analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology* 9 (2002).