FINANCIAL TIMES

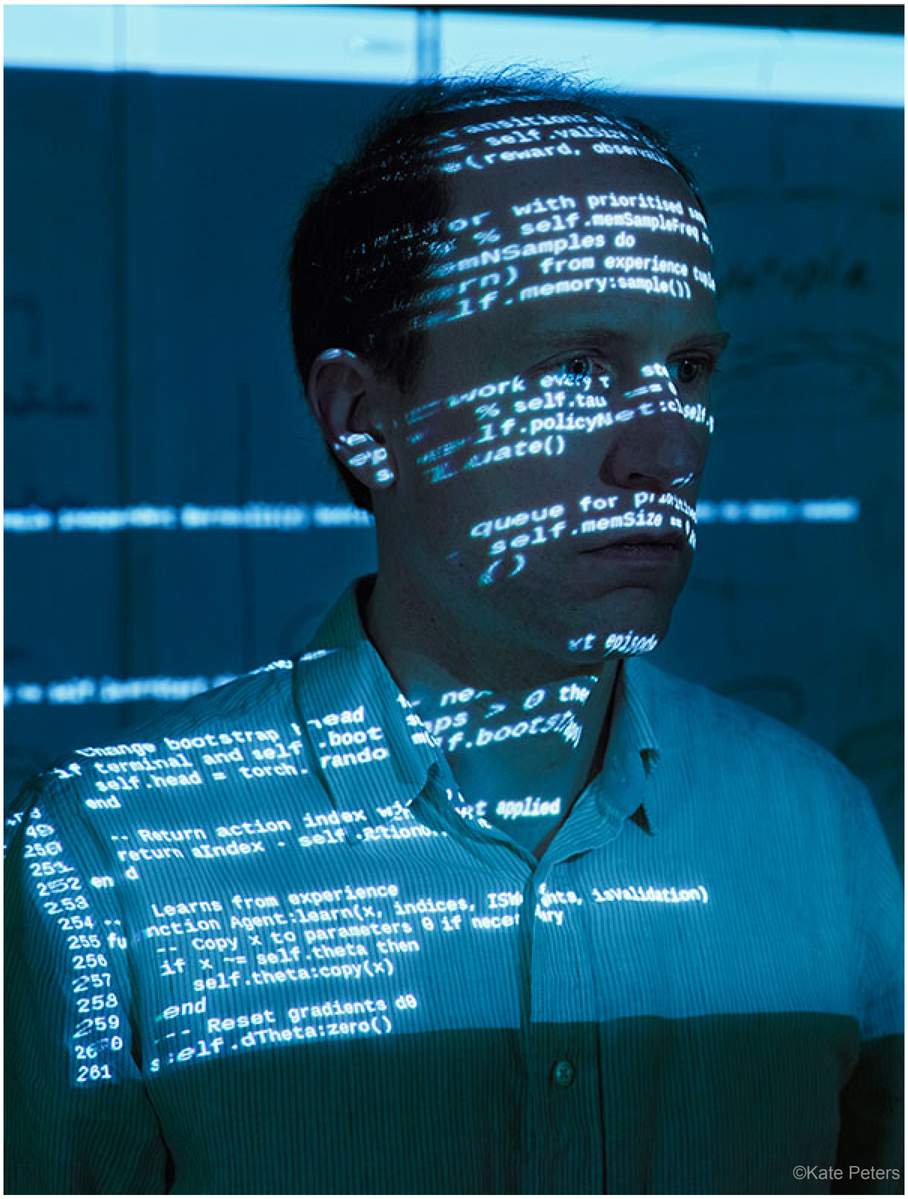
July 14, 2016 7:29 am

Artificial intelligence: can we control it?

John Thornhill

Share Print Clip Comments

It is the world's greatest opportunity, and its greatest threat, believes Oxford philosopher Nick Bostrom



©Kate Peters

Professor Nick Bostrom with code from DeepMind's DQN program that could teach itself to play computer games at or above human level

Scientists reckon there have been at least five mass extinction events in the history of our planet, when a catastrophically high number of species were wiped out in a relatively short period of time. We are possibly now living through a sixth — caused by human activity. But could humans themselves be next?

This is the sort of question that preoccupies the staff at the Future of Humanity Institute in Oxford. The centre is an offshoot of the Oxford Martin School, founded in 2005 by the late James Martin, a technology author and entrepreneur who was among the university's biggest donors. As history has hit the fast-forward button, it seems to have become the fashion among philanthropists to endow research institutes that focus on the existential challenges of our age, and this is one of the most remarkable.

Tucked away behind Oxford's modern art museum, the institute is in a bland modern office block; the kind of place you might expect to find a provincial law firm. The dozen or so mathematicians, philosophers, computer scientists and engineers who congregate here spend their days thinking about how to avert catastrophes: meteor strikes, nuclear winter, environmental destruction, extraterrestrial threats. On the afternoon I visit there is a fascinating and (to me) largely unfathomable seminar on the mathematical probability of alien life.

Presiding over this extraordinary institute since its foundation has been Professor Nick Bostrom, who, in his tortoise-shell glasses and grey, herringbone jacket, appears a rather ordinary academic, even if his purple socks betray a streak of flamboyance. His office resembles an Ikea showroom, brilliantly lit by an array of floor lamps, somewhat redundant on a glorious sunny day. He talks in a kind of verbal origami, folding down the edges of his arguments with precision before revealing his final, startling conclusions. The slightest monotonal accent betrays his Swedish origins.

. . .

Bostrom makes it clear that he and his staff are not interested in everyday disasters; they deal only with the big stuff: "There are a lot of things that can go and have gone wrong throughout history — earthquakes and wars and plagues and whatnot. But there is one kind of thing that has not ever gone wrong; we have never, so far, permanently destroyed the entire future."

Anticipating the obvious next question, Bostrom argues that it is fully justified to devote resources to studying such threats because, even if they are remote, the downside is so terrible. Staving off future catastrophes (assuming that is possible) would bring far more benefit to far greater numbers of people than solving present-day problems such as cancer or extreme poverty. The number of lives saved in the future would be many times greater, particularly if "Earth civilisation", as he calls it, spreads to other stars and galaxies. "We have a particular interest in future technologies that might potentially transform the human condition in some fundamental way," he says.

So what tops the institute's list of existential threats? A man-made one: that rapidly advancing research into artificial intelligence might lead to a runaway "superintelligence" which could threaten our survival. The 43-year-old philosopher is himself a notable expert on AI and the author of *Superintelligence* , a startling and controversial book that discusses what he describes as "quite possibly the most important and most daunting challenge humanity has ever faced."

"Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct," he wrote.

Dealing with the big stuff: staff at the Future of Humanity Institute in Oxford

Published in 2014, the book caused an immediate stir in academic circles and, more surprisingly, climbed into The New York Times bestseller list. Bill Gates, the intellectually omnivorous co-founder of Microsoft, strongly recommended it.

Some AI experts have accused Bostrom of alarmism, suggesting that we remain several breakthroughs short of ever making a machine that "thinks", let alone surpasses human intelligence. A sceptical fellow academic at Oxford, who has worked with Bostrom but doesn't want to be publicly critical of his work, says: "If I were ranking the existential threats facing us, then runaway 'superintelligence' would not even be in the top 10. It is a second half of the 21st century problem."

But other leading scientists and tech entrepreneurs have echoed Bostrom's concerns. Britain's most famous scientist, Stephen Hawking, whose synthetic voice is facilitated by a basic form of AI, has been among the most strident. "The development of full artificial intelligence could spell the end of the human race," he told the BBC.

Elon Musk, the billionaire entrepreneur behind Tesla Motors and an active investor in AI research, tweeted: "Worth reading *Superintelligence* by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes."

AI is not just a cool gadget or a nifty little thing. it's the last invention humans will ever need to make

- Nick Bostrom

Although Bostrom has a reputation as an AI doomster, he starts our discussion by emphasising the extraordinary promise of machine intelligence, in both the short and long term. "I'm very excited about AI and I think it would be a tragedy if this kind of superintelligence were never developed." He says his main aim, both modest and messianic, is to help ensure that this epochal transition goes smoothly, given that humankind only has one chance to get it right.

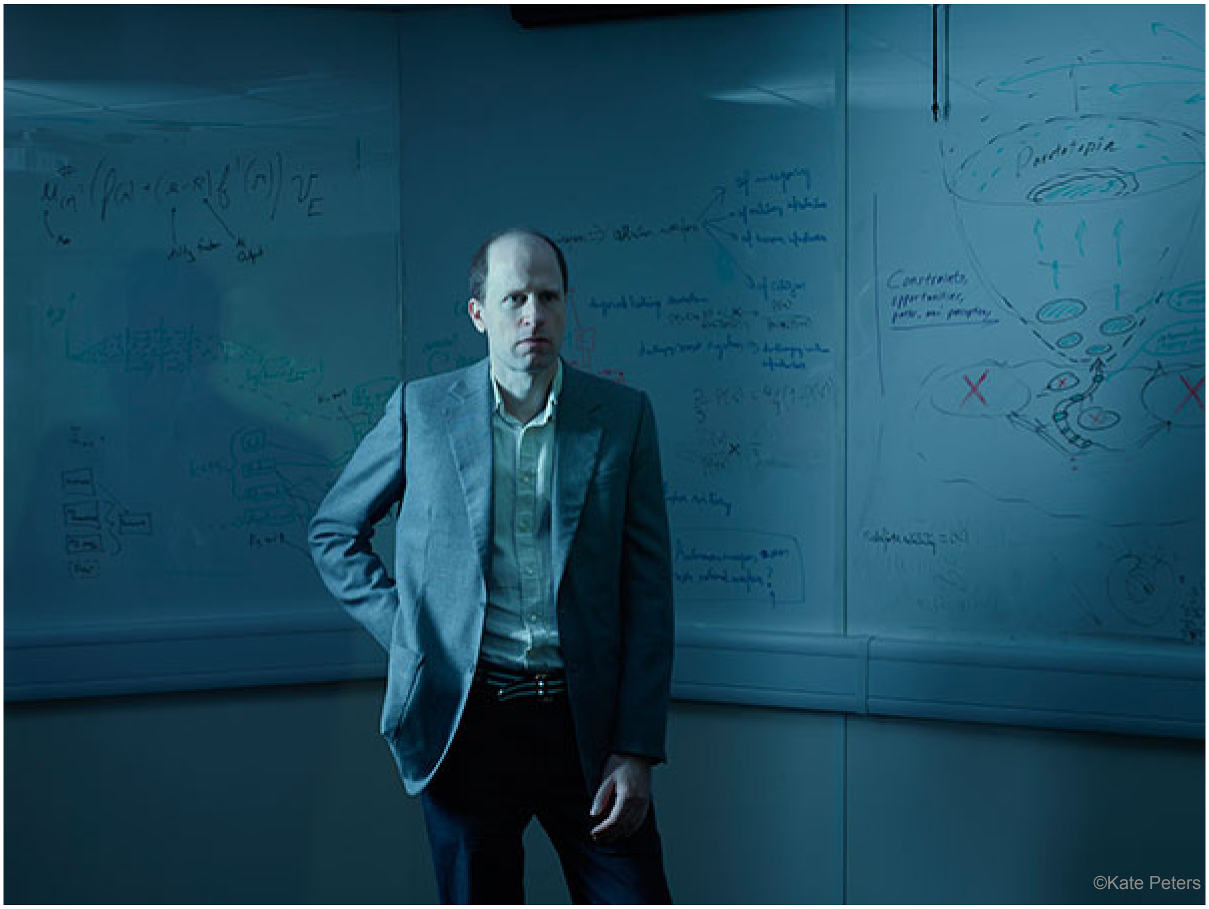
"So much is at stake that it's really worth doing everything we can to maximise the chances of a good outcome," he says. "It's not just another cool gadget or another nifty little thing. It's the last invention humans will ever need to make." In the industrial revolution, he explains, we automated a lot of physical labour to develop artificial muscle. "With the AI transition we will automate human thoughts, human brain power. It's hard to think of any important area of human life that would not be impacted in some way."

If we do ever reach the stage when computers can design computers, then AI could accelerate the process of discovery, he argues. All inventions that might have been made by humans over the next 40,000 years could be compressed into the immediate future, opening up the possibility of colonising space or stopping human ageing.

One of the biggest unknowable factors is how quickly we may develop Human Level Machine Intelligence (HLMI) — if ever. In *Superintelligence*, Bostrom cited a survey of many of the world's top AI researchers, conducted by Nils Nilsson, a leading American computer scientist. Based on this survey, Nilsson estimated there was a 10 per cent chance of reaching HLMI by 2030, a 50 per cent chance by 2050, and 90 per cent by 2100.

However, Stuart Russell, professor of computer science at University of California, Berkeley, says the scientific community's ability to predict breakthroughs is often limited, citing the example of nuclear power. In September 1933, the eminent physicist Ernest Rutherford gave a speech in which he said that anyone who predicted energy could be derived from the transformation of the atom "was talking moonshine". The very next day the Hungarian scientist Leo Szilard worked out conceptually how it could be done.

"The gap between authoritative statements of technological impossibility and the 'miracle of understanding' . . . that renders the impossible possible may sometimes be measured not in centuries . . . but in hours," Russell wrote in an online piece for the website edge.org.


©Kate Peters

Bostrom worries that AI research may be too open, rather than too closed

Bostrom himself is wary of making any predictions about the immediate future. It is possible, he says, that someone might break through a conceptual "false wall", opening up a clear field in front of us and enabling rapid breakthroughs. Equally, it is possible for researchers to become bogged down. We have, after all, lived through previous AI "winters", when startling advances were followed by slow progress.

The human ability to strategise and plan in the face of changing circumstances is fiendishly difficult to replicate. However, the recent success of AlphaGo, a programme created by Google DeepMind, in defeating the world champion at the ancient game of Go suggests that we are moving faster than previously thought. When *Superintelligence* was published, the prediction was that it would take AI a decade to beat the best human player. Instead, AlphaGo accomplished the task this year. "Progress in AI has been faster over the past two years than most people expected," Bostrom says.

A huge amount of money is being poured into AI by tech companies, principally in the US and China. Google, which acquired the London-based DeepMind in 2014, has made AI one of its priorities. Facebook, Baidu, Microsoft, and IBM are also investing heavily in research and applying machine-learning programmes to their commercial activities. Such is the demand that the annual salary for a post-doctoral AI specialist in Silicon Valley is about $400,000 a year.

Some academics worry that too much of the latest cutting-edge research into AI is being conducted by private corporations, rather than public universities. Should such potentially mind-blowing technologies be in the hands of these rich and powerful entities, more concerned with commercial opportunity than disseminating knowledge?

AlphaGo v Lee Sedol: In 2014, scientists thought it would take a decade for a machine to beat the world champion at Go. The AlphaGo programme won in March this year.

Bostrom, however, worries that the latest AI research may be too open, rather than too closed. One of the most interesting recent developments in the AI debate was the creation last year of OpenAI, backed by Elon Musk and other tech luminaries such as Peter Thiel and Reid Hoffman. Its stated aim is to create a non-profit research company that will "benefit humanity as a whole". OpenAI will encourage its researchers to publish their work, share any patents they develop and collaborate freely with outside institutions.

But Bostrom is not convinced that this commitment to the free flow of information is such a self-evident virtue. He recently wrote a paper about the dangers of too much transparency in publishing research, given that AI could be used for harmful purposes as well as good. OpenAI's main strength, in Bostrom's view, is that it is a non-profit organisation with the goal of promoting AI as a universal benefit. If we ever attain general AI, he says, there will be plenty of advantages for everyone. ("It would be like Christmas 365 days a year.") Yet he remains concerned that less scrupulous researchers, rushing to develop HLMI ahead of others, might ignore sensible safety mechanisms or even look to use it for destructive ends.

"I don't know anybody who advocates that software for lethal autonomous weapons should be open-sourced," he points out. He also worries about dictators getting their hands on the latest surveillance, data-mining, and predictive techniques. "Is it good to enable potential despots to protect themselves against possible insurrection and conspiracy by making this technology available?" he asks.

. . .

Born in Helsingborg in Sweden, Bostrom claims to have been bored at school until he had an "intellectual awakening" in his mid-teens. Resolving to make up for lost time, he studied psychology, philosophy, physics and mathematical logic at university before completing a PhD at the London School of Economics. Somewhat improbably, given his serious mien, he also found time to do some turns on London's stand-up comedy circuit. His intellectual interests range widely and he has published more than 200 papers on a variety of topics. He is married to a doctor who lives in Vancouver with their young child, and conducts much of his personal life via Skype.

Bostrom says he was both surprised by the reaction of AI experts to his book and heartened by their constructive response. The public conversation has come a long way since even a few years ago, he says, when it mainly consisted of science fiction tales about robots taking over the world.

A debate is already raging about the impact that AI could have on privacy and employment. Machine-learning can now be used to compile incredibly detailed digital profiles of individuals by meshing together social media profiles, browsing histories and location tracking, prompting regulators to consider new privacy protections. The widespread use of AI — in driverless cars, for example — could also destroy thousands of jobs, even if it may eventually help create employment in areas we cannot imagine today.

But there's also a growing discussion about the "control problem", as Bostrom calls it. How do we ensure that computers do not acquire a mind of their own, threatening human existence?

In a short story written in 1942, Isaac Asimov introduced his "Three Laws of Robotics", which have served as a very rough frame of reference for the debate ever since. Asimov later added a fourth, or "zeroth", law that trumped all others in importance: "A robot may not harm humanity [as a whole], or, by inaction, allow humanity to come to harm."

But trying to ensure such laws are followed in practice is a phenomenal challenge. Bostrom compares the debate about AI to similar agonising among physicists when developing the nuclear bomb, and among biologists when discussing gene editing. If scientists cannot "unlearn" knowledge and prevent discoveries being made, they should at least help frame the debate so that these technologies are used responsibly. One good example came in 1975 when a group of scientists met at Asilomar in California to establish a set of principles for regulating biotechnology.

In relation to AI, he says, "The helpful way to engage with this issue is to try to make sure that it is done, but it is done right." In particular, he argues that it's essential to lay the foundations for "building up the technology and the science of understanding how to predict and control advanced artificial agents". It's also vital that all sides in the debate work together rather than "lobbing rhetorical grenades over a wall".

> Before the prospect of AI we are like small children playing with a bomb
>
> - Nick Bostrom

Bostrom stresses again that he is not hostile to the development of new technology, as he has sometimes been portrayed. Indeed, he has been a long-time supporter of the transhumanist movement, which seeks to transform the human condition by using technology to enhance our intellectual and physical capabilities. And as a convinced transhumanist, he is massively in favour of new breakthroughs. “Bring me my life extension pill right away. Pump in the cognitive enhancing drug,” he says with a thin smile.

When *Superintelligence* was published, criticisms tended to fall into two camps. There were those who dismissed Bostrom’s analysis as trivial or tendentious, suggesting that the “control problem” could be easily solved or may never arise. Then there were those who said it would be impossible to control a superintelligence, so there was no point in even trying. Their common position, according to Bostrom, was that neither camp felt the need to make any effort to address the issue today.

Bostrom, however, believes there is a third possibility: that the problem may be difficult, but is not insoluble, provided we start early enough and apply enough mathematical talent. What would be terrible, in his view, would be to find ourselves on the brink of developing HLMI and realising that it’s too late to do anything to ensure humans retain control over our creations. “That seems like a stupid place to be, if we can avoid it.”

“Maybe the problem will turn out to be much easier than it seems, and that will be good. It still seems extremely prudent, though, to put in the work, in case it does turn out to be harder,” he says. “Whether the risk is 1 per cent, 80 per cent, or anywhere in between, it still makes sense to do some of the things I think should be done. The position is not sensitive to one’s level of optimism.”

At this stage, he doesn’t believe governments need to be much involved. “It would be premature today, I think, to try to introduce some regulation.” But there are areas where AI is being applied — such as self-driving cars, data privacy, and surveillance — where government regulation could play a useful role in shaping the industry’s evolution. He believes there should be more public discussion about the practical ethical issues surrounding the introduction of all new technologies.

Bostrom says that, partly thanks to his intervention, the public debate about AI is developing in a healthy way. The interest is there; it just needs to be channelled in a constructive direction. He and his colleagues at the Future of Humanity Institute are trying to devise conceptual strategies to help address the control problem. “I don’t think just ringing the bell and screaming is actually what the world needs,” he says. “It needs to get on with the actual hard work.”

John Thornhill is the FT’s innovation editor

Photographs: Kate Peters

RELATED TOPICS Artificial Intelligence and Robotics

Share Print Clip Comments

VIDEO

Printed from: http://www.ft.com/cms/s/0/46d12e7c-4948-11e6-b387-64ab0a67014c.html