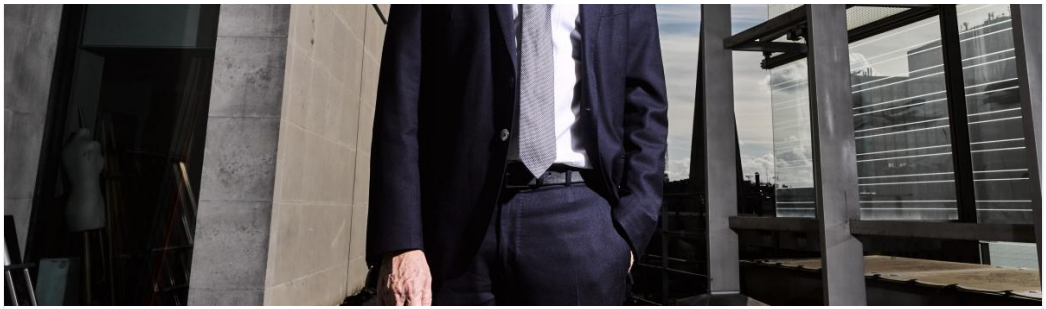
INTERVIEW

Is the rise of killer machines closer than we think?

British computer scientist and world-leading AI expert Stuart Russell says we urgently need to understand the potential of superintelligent machines – before they turn on us. By Damian Whitworth

‹ PREVIOUS ARTICLE NEXT ARTICLE ›



Stuart Russell, 59
RICHARD ANSETT/BBC

Damian Whitworth
Saturday January 29 2022, 12.01am GMT, The Times

A couple of years ago Stuart Russell, a British computer scientist who is one of the world's leading experts on artificial intelligence, was approached by a film director who wanted him to be a consultant on a movie. The director complained that there was too much doom and gloom about the future of superintelligent machines. He wanted Russell to explain how the human heroes in the film could save our species by outwitting AI. "Sorry," Russell told the director. "They can't."

Russell is a professor of computer science at the University of

PREVIOUS ARTICLE NEXT ARTICLE

and the White House, and co-written the standard university textbook on artificial intelligence. Success in creating superintelligent AI, he has predicted, “would be the biggest event in human history... and perhaps the last event in human history”.

AI could lead us into a golden age, where we can enjoy lives that are no longer burdened by drudgery. Or it could destroy us as a species. Even if we learn to live with superintelligent machines, they may take all our jobs or create mayhem on battlefields. Vladimir Putin has said whoever takes the lead in AI “will become the ruler of the world”, prompting the billionaire entrepreneur Elon Musk to predict that nations competing for AI superiority will be the most likely cause of a third world war.

When Russell gave the Reith lectures last year, the headlines were mostly about the havoc that lethal autonomous weapons systems could wreak. But Russell has a wider vision, which is by turns thrilling and more terrifying than coronaviruses and global warming.

While the human brain has evolved over millions of years, the development of computers and robots to simulate the human mind’s ability to solve problems, make decisions and learn has taken a few decades. From the very beginning of AI, says

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

He believes we should make a very significant tweak to that definition so that machines are seen as “beneficial” to the extent that their actions can be expected to achieve “our” objectives. If we don’t design them with our wellbeing specifically in mind, we could be creating an existential problem for ourselves.

In the past decade AI has started to fulfil some of its promise. Machines can thrash us at chess. When Russell was taking a sabbatical in Paris, he used machine translation to complete his tax return. In a recent breakthrough that could transform

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

phone that has learnt to recognise my voice and provides a reasonable simultaneous transcription of our conversation (although its claim, for example, that Russell is talking about "kick-ass machines made of cheese" does underline that AI armageddon is still some way off).

These AIs are limited to harnessing considerable computational power to complete well-defined tasks. Google's search engine "remembers" everything, but can't plan its way out of a paper bag, as Russell puts it. The goal of AI research is creating a general-purpose AI that can learn how to perform the whole range of human tasks from, say, teaching to running a country. Such a machine "could quickly learn to do anything that human beings can do", says Russell. And given that computers can already add billion-digit numbers in a fraction of a second, "Almost certainly it would be able to do things that humans can't do."

The creation of a superintelligent AI, which Russell has likened to the arrival of a superior alien civilisation (but more likely), is an enormous challenge and a long way off. But many experts believe it could happen in the next few decades, and Russell is an evangelist for the need to prepare for such an eventuality.

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

Three bosses

He likes to talk about Alan Turing, the father of theoretical computer science and AI, who in 1951 gave a lecture in which he chillingly predicted the arrival of superintelligent machines. “It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers,” said Turing. “At some stage therefore we should have to expect the machines to take control.”

The danger, Russell suggests, is that our relationship with machines becomes analogous to the relationship gorillas have with us today. We had a common ancestor but “once humans came along, and they’re this much more intelligent than gorillas and chimpanzees, then game over. I think that’s sort of how Turing saw it. Intelligence is power. Power is control. That will be the end of it.”

Russell doesn’t believe that is necessarily the end of it, if we go about things the right way. But he wants us to be clear about the threat. Science fiction has sometimes suggested that machines will supersede us when they develop human consciousness; that when they are aware of themselves and their surroundings

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

their advanced competency. A highly sophisticated machine with a fixed objective could stop at nothing to achieve that objective and fail to take into account other human priorities.

He calls this the “King Midas problem” after the mythical figure who asked for everything he touched to be turned to gold, realising too late that this would include food, drink and his family.

Already we give machines objectives that are not perfectly aligned with our own. Social-media algorithms are designed to maximise click-through in order to keep people on the site and

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

algorithm works out what keeps them online and the diet of content they are fed is contributing to growing extremism around the world. “When a person is interacting with a system for six or eight hours a day, the algorithm is making choices that affect your behaviour, nudging you hundreds of times a day. And that’s happening to billions of people.” He would love to see the internal data from big tech companies “to really understand what’s going on”, but adds, “In America, you’ve got 60 million people who are living in a fantasy world.”

Imagine a more sophisticated AI that is capable of going into a coffee shop to get you a latte. It will be unhelpful to café society if it tears the place apart because it is fixed on achieving the task whatever the cost.

Here we are entering the territory of *2001: A Space Odyssey*, in which Hal, the spaceship computer, kills four of the five astronauts on board because he deems them a threat to the mission.

AI’s potential to help in medicine is already being realised, but Russell raises the spectre of a superintelligent AI system being charged with finding a cure for cancer. It could quickly digest all the literature and make hypotheses, but all that will be wildly counterproductive if it then concludes that the quickest way to

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

We might recruit an AI to fight the acidification of the oceans, only to find that its solution is to use a quarter of the oxygen in the atmosphere to achieve this and we all asphyxiate.

Solving the King Midas problem also solves the gorilla problem, by ensuring that AI is not in conflict with humans and we don't end up existing at the whim of the machines.

So we need to create AI systems carefully. They must be built so they are altruistic towards humans and uncertain about what all our preferences are. Then the AI system would ask what our preferences are regarding oxygen before going ahead and deacidifying the oceans.

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

there may be other things we care about. So if we say, 'I'd like a cup of tea,' that doesn't mean you can mow down all the other people at Starbucks to get to the front of the line."

And the machine must be devised so it will always allow us to turn it off. Otherwise, its logical conclusion would be to deactivate its "off" switch in order to eliminate an obvious threat to completing the task.

Given the starkness of some of his misgivings about the future, I was expecting Russell to be an intense prophet of cyber-doom in real life, but he is reasonable, softly spoken with a mid-Atlantic accent, and often funny, displaying an understated wit that is familiar from some of his writings.

He is in London for a holiday with his wife, Loy Sheflott, founder and CEO of Consumer Financial, a marketing firm for financial services companies. They have four children who range in age from 15 to 23.

Russell, 59, was born in Portsmouth and moved around the country because of his father's job running Crown Paints and Wallcoverings. They also lived in Toronto for a few years. His mother was a fashion designer and teacher. Russell boarded at St Paul's School in southwest London where even in an

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

could study the subject for A-level.

He left school at 16 having taken his A-levels early, spent a gap year at IBM and then, at 17, went to Oxford, where he was awarded a first in physics. He moved to the US to do a PhD in computer science at Stanford University and then joined the University of California at Berkeley, where he is professor of electrical engineering and computer sciences and director of the Centre for Human-Compatible Artificial Intelligence. With Peter Norvig, Google's former research director, he wrote the standard university textbook on AI and in his most recent book, *Human Compatible: AI and the Problem of Control*, he outlined some of his concerns about the future of artificial intelligence.

Even if machines don't take over the planet and eradicate us and we find a way to stay in control, living with them may present enormous challenges. What happens when they can do all – or, at least, the vast majority – of the roles that fill our working days? While he says they are currently useless at interviewing, it seems a reasonable bet that there are future interviewers being born today who will be made redundant by AI, along with house painters, drivers and radiographers.

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

For many millennia, Russell points out, most humans have been in “robot” jobs; if they are released from agricultural, industrial and clerical roles by real robots it could transform human existence. “If all goes well, it will herald a golden age for humanity. Our civilisation is the result of our intelligence, and having access to much greater intelligence could enable a much better civilisation,” he said in one of his Reith lectures.

Robots could build bridges, improve crop yields, cook for 100 people, run elections, while we get on with... what? We would

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

A lot of us, suggests Russell, will be engaged in interpersonal services, supplying our humanity to others, whether as therapists, tutors or companions. We would have all the time in the world to strive to perfect the art of living, through art, gardening or playing games.

“The need will not be to eat or be able to afford a place to live, but the need for purpose,” says Russell. We are used to adapting to new jobs, but less so to having no job at all.

Is there not a danger that we end up with millions of therapists and slightly crap artists? “I don’t feel that’s the route to fulfilment,” he says, smiling.

The most immediate problem facing us comes in the form of lethal autonomous weapons. They are already with us. The threat is not that AI weapons are going to turn upon us because our objectives and theirs collide, but that they can be used by nefarious states or groups to target their enemies.

Israel’s Harop has a 10ft wingspan and the ability to loiter and search for targets and, when it recognises them, make a kamikaze attack. The UN has reported that a smaller drone may have autonomously targeted militia fighters in Libya.

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

and then track people through technology that recognises a face or “anything you want: yarmulkes or turbans or whatever”.

He can envisage a mass attack by a swarm. “I think it could happen that we would get attacks with a million weapons.”

We’ve legislated internationally against biological and chemical weapons and to stop nuclear proliferation. The systems are not perfect, but do mean the world community can go after those who don’t comply and make it hard for them to get the ingredients to create these weapons. Russell is frustrated by the reluctance of governments, including the UK and US, to ban lethal autonomous weapons outright. Officials at the Obama White House listened very carefully when he was part of a delegation there. “Their response on weapons of mass destruction was, ‘But we would never make weapons like that.’ In that case, why won’t you ban them? And they didn’t have an answer.”

I joke that by now computers must all know who he is and are probably listening in on this conversation and swapping notes. “I’m just trying to prevent the machines from making a terrible mistake,” he says.

A small part of me is paranoid that someone – or some artificial

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

cotton bud in what I thought was the privacy of my own home office? I can't believe I'm telling Russell this, but I keep a sticky note over the lens when I'm not on a video call. Rather to my surprise, he says, "I think that's a good idea." People who know more about computer security than he does say the same apparently.

I wonder what he thinks of Elon Musk's hopes to build a brain-machine interface or "neural lace", inspired by Iain M Banks's Culture novels. "His solution to the existential risk is that we actually merge with the machines," he says. "If we all have to have brain surgery just to survive, perhaps we made a mistake somewhere along the line."

How worried is he that his children or any future grandchildren will face a dystopian future with AI? "It doesn't feel like a visceral fear. It feels like climate change." But in the worst-case scenario AI would be terminal for our species, whereas with climate change we could probably cling on in the last temperate corners of the world. So AI could be worse than global warming? "In the worst case, yes. We have to follow our reasoning where it leads us. And if the machines really are more intelligent than us and we've made a mistake and set them up to pursue objectives that end up having these disastrous side effects, we would have no more power than chess players have when they are playing

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

that?" he says, raising an eyebrow. On the way over on the plane he was playing a rather more formidable chess programme. "It doesn't let you take any moves back."

Sony's Aibo, a robotic puppy

AI: the next 10 years

By Monique Rivalland

Health

The race is on to transform healthcare with AI and the market is estimated to be worth £120 billion by 2028. So what can we

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

can already identify signs of diabetic retinopathy from eye scans with 90 per cent accuracy. At hospitals and care homes basic nursing tasks could be carried out by AI assistants. The field of neuroprosthetics, which develops brain implants, robotic limbs and cyborg devices, will help us overcome cognitive and physical limitations. This month BioNTech, maker of the Pfizer Covid-19 vaccine, launched an "early warning system" with London-based AI firm InstaDeep to detect new variants of the coronavirus before they spread.

Pets

Japan is leading the way in AI pets. Sony's Aibo, which costs £2,127, is a robotic puppy. Aibo will respond to commands as well as read human emotions and distinguish between family members. When tired, Aibo returns to his charging station. Towards the end of 2020, almost a year into the pandemic, local government in New York started offering AI-powered furry tabby cats from robotics company Joy For All to care homes and older people in social isolation. China's Unitree wants to make its four-legged robots, currently £1,980, as affordable as phones. It won't be long before AI companions need not resemble traditional pets for humans to warm to them. Spot The Dog is not exactly a pet but a robotic canine that is so agile it is used to explore remote environments too dangerous or extreme for humans. Made by Boston Dynamics and sold for £55,312, it

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

Robots and drones could carry out perilous tasks such as bomb disposal, but the biggest change to warfare will come in the shape of artificially intelligent killing machines. In November 2020, Israel assassinated Iran's top nuclear scientist using a high-tech, computer-powered sharpshooter with multiple camera eyes, capable of firing 600 rounds a minute.

Transport

There are more than ten unicorn start-ups – that's companies valued at $1 billion – vying for leadership in the autonomous vehicle industry. They're in China, America, Britain and Canada and include personal transport as well as trucks and haulage. This month the MK Dons (Milton Keynes) football team have been trialling driverless cars called Fetch to take them to and from training. Self-driving cars are supposed to be safer and more efficient than human drivers and are expected on British roads later this year. The government has announced that cars fitted with automatic lane-keeping systems will be permitted to drive at up to 37mph in a single lane without the driver interacting with it.

Education

The main benefit here is that AI will better tailor education to students' needs. Virtual tutors will assist human teachers in the classroom, offering support to students by giving instant

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

Communication
Microsoft and Skype already have a voice translator that can translate between 11 languages, including Chinese, English, French, Japanese, Russian and Spanish. This is likely to advance quickly to real-time translation of hundreds of languages, taking us a step closer to universal conversation. Google is working on an AI assistant that can complete simple phone-based tasks such as calling your doctor to make an appointment. No more waiting on hold.

Media
Journalists, beware. Simple or factual news will increasingly be written by algorithms. It has started: *The Washington Post*'s "AI Writer" wrote more than 850 stories during the Rio Olympics in 2016; Bloomberg uses AI tech to relay complex data, and Associated Press uses natural language AI to produce 3,700 earnings reports a year.

Sources *Forbes, McKinsey, Statists, IHS Markit, Pega*

Related articles

COMMENT | CRISTIN LEACH
We should be wary of the tasks we offload to artificial

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

December 11 2021, 12.01am GMT

Cristín Leach

AI software helps mathematicians pinpoint patterns

"A mathematician," the Hungarian number theorist Alfred Renyi said, "is a machine for turning coffee into theorems". The...

December 02 2021, 12.01am GMT

Tom Whipple, Science Editor

INTERVIEW

How to save the world — by a man who might actually know

Mo Gawdat glimpsed the apocalypse in a robot arm. Or rather, in a bunch of robot arms, all being developed together. An arm...

September 29 2021, 12.01am BST

Hugo Rifkind

Comments are subject to our community guidelines, which can be viewed here.

Stuart Russell

S Add to the conversation...

Sort by **Recommended**

H **Helios** • 9 HOURS AGO

'AI' has become a buzzword used far too liberally. Most so called 'AI' has no 'intelligence' but is merely a sophisticate dumb algorithm.

Reply ☆ Recommend (6)

S **Seldon** • 9 HOURS AGO

Yup. Linear regression about 90% of the time when used in corporate speak.

Reply ☆ Recommend (5)

L **LarryC** • 2 HOURS AGO

Machine Learning is a much more common term these days than AI: perhaps a concession that the processes aren't that smart, but they are effective. It's quite sobering to consider that the approach of defining an objective (but not necessarily the means of obtaining it) is resulting in near-human le...**See more**

Mike S • 9 HOURS AGO

Fascinating. Thank you for this article

Reply ☆ Recommend (5)

Glasgow Kiss • 9 HOURS AGO

The Singularity has been reached, we just don't know it yet.

Reply ☆ Recommend (1)

Mospe • 7 HOURS AGO

Amongst human beings it is not the most intelligent who end up being in charge, witness are politicians. We need to concentrate on producing biddable AI Like Musk is producing a rather weedy robot

Reply ☆ Recommend

MH18 • 1 HOUR AGO

AI possibilities are constantly referred to as capabilities. You only have to try to get a chatbot to understand and answer a relatively simple question to realise how absolutely bloody useless they are. Referring you to a human is not an example AI solving a problem.

Reply ☆ Recommend

Howard Jones • 2 HOURS AGO

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

What none of these mode...**See more**

Reply ☆ Recommend (1)

2022RAM • 7 HOURS AGO

Has God been using AI to control the universe for eons? God knows.

Reply ☆ Recommend

Johnny H • 4 HOURS AGO

I'd love to know what he thinks about Isaac Asimov's 3 Laws of Robotics.
1. A robot shall not harm a human, or by inaction allow a human to come to harm.
2. A robot shall obey any instruction given to it by a human.
3. A robot shall avoid actions or situations that could cause it to come to harm its...**See more** *(Edited)*

Reply ☆ Recommend

Bat-Ori • 8 HOURS AGO

'Alexa, wash all my blue shirts except the short-sleeved ones and the one I wore yesterday'. 'Sorry, I don't know that song'. Are battlefield robots more competent than that?

Reply ☆ Recommend

‹ PREVIOUS ARTICLE NEXT ARTICLE ›

Reply ☆ Recommend

View more comments

OpenWeb Feedback

BACK TO TOP

GET IN TOUCH

About us

Contact us

Help

The Times Editorial Complaint

The Sunday Times Editorial Complaints

Place an announcement

Classified advertising

Display advertising

The Times corrections

The Sunday Times corrections

Careers

MORE FROM THE TIMES AND THE SUNDAY TIMES

The Times e-paper

Times Currency Services

Times Crossword Club

Times+

Times Expert Traveller

Schools Guide

Best Places to Live

Sportswomen of the Year Awards

Podcasts

Privacy & cookie policy

Cookie settings

Topics

Terms and conditions

Licensing

Site map

Commissioning terms

Do not sell my personal information

‹ PREVIOUS ARTICLE

NEXT ARTICLE ›