The end of humanity: will artificial intelligence free us, enslave us — or exterminate us?

The Berkeley professor Stuart Russell tells Danny Fortson why we are at a dangerous crossroads in our development of AI

Scenes from Stuart Russell's dystopian film Slaughterbots, in which armed microdrones use facial recognition to identify their targets

Danny Fortson
Saturday October 26 2019, 11.01pm GMT, The Sunday Times
Share

Stuart Russell has a rule. “I won’t do an interview until you agree not to put a Terminator on it,” says the renowned British computer scientist, sitting in a spare room at his home in Berkeley, California. “The media is very fond of putting a Terminator on anything to do with artificial intelligence.”

The request is a tad ironic. Russell, after all, was the man behind Slaughterbots, a dystopian short film he released in 2017 with the Future of Life Institute. It depicts swarms of autonomous mini-drones — small enough to fit in the palm of your hand and armed with a lethal explosive charge — hunting down student protesters, congressmen, anyone really, and exploding in their faces. It wasn’t exactly Arnold Schwarzenegger blowing people away — but he would have been proud.

Autonomous weapons are, Russell says breezily, “much more dangerous than nuclear weapons”. And they are possible today. The Swiss defence department built its very own “slaughterbot” after it saw the film, Russell says, just to see if it could. “The fact that you can launch them by the million, even if there’s only two guys in a truck, that’s a real problem, because it’s a weapon of mass destruction. I think most humans would agree that we shouldn’t make machines that can decide to kill people.”

The 57-year-old from Portsmouth does this a lot: deliver an alarming warning about the existential threat posed by artificial intelligence (AI), but through a placid smile. “We have to face the fact that we are planning to make entities that are far more powerful than humans,” he says. “Howdo we ensure that they never, ever have power over us?” I almost expect him to offer a cup of tea to wash down the sense of imminent doom.

There is no shortage of AI doom-mongers. Elon Musk claims we are “summoning the demon”. Stephen Hawking famously warned that AI could “spell the end of the human race”. Seemingly every month, a new report predicts mass unemployment and social unrest as machines replace humans.

The bad news? Russell, essentially, agrees with all of it. This is disconcerting because he quite literally wrote the book on the technology. His textbook, Artificial Intelligence: A Modern Approach, is the most widely used in the industry. Since he authored it in 1994 with Peter Norvig, Google’s director of research, it has been used to train millions of students in more than 1,000 universities.

Now? The University of California, Berkeley professor is penning a new edition where he admits that they “got it all wrong”. He adds: “We’re sort of in a bus and the bus is going fast, and no one has any plans to stop.” Where’s the bus going? “Off the cliff.”

The good news, though, is that we can turn the bus around. All it entails is a fundamental overhaul, not only of how this frighteningly powerful technology is conceived and engineered, but also of how we, as a corpus of nearly 8bn people, organise, value and educate ourselves.

From Russell’s vantage point, we have come to a crossroads. In one direction lies “a golden age of humanity” where we are freed from drudgery by machines. The other direction is, well, darker. In his new book, called Human Compatible, Russell sums it up with what he calls “the gorilla problem”.

Apes, our genetic progenitors, were eventually superseded. And now? “Their species has essentially no future beyond that which we deign to allow,” Russell says. “We do not want to be in a similar situation vis-à-vis super-intelligent machines.” Quite.

Russell came to California in the 1980s to get a PhD after Oxford, and never left. He is an insider but with an outsider's perspective. Talk to most computer scientists and they scoff at the idea that has him so worried: artificial general intelligence, or AGI. It is an important distinction. Most of the AI out in the world today involves what is known as "machine learning".

These are algorithms that crunch through inconceivably large volumes of data, draw out patterns, then use those patterns to make predictions. Unlike previous AI booms (and busts), dramatic reductions in the cost of data storage coupled with leaps in processing capability mean that algorithms finally have enough horsepower and raw data to train on. The result is a blossoming of suddenly competent tools that are also sometimes wildly powerful. They are, however, usually designed for very defined, limited tasks.

Take, for example, a contest organised by several American universities last year between five experienced lawyers and an AI designed to read contracts. The goal was to see who was better at picking out the loopholes. It was not a great day for Homo sapiens. The AI was not only more accurate — it found 94% of the offending passages, the humans uncovered 85% — but it was faster. A lot faster. While the lawyers needed an average of 92 minutes to complete the task, the AI did it in 26 seconds.

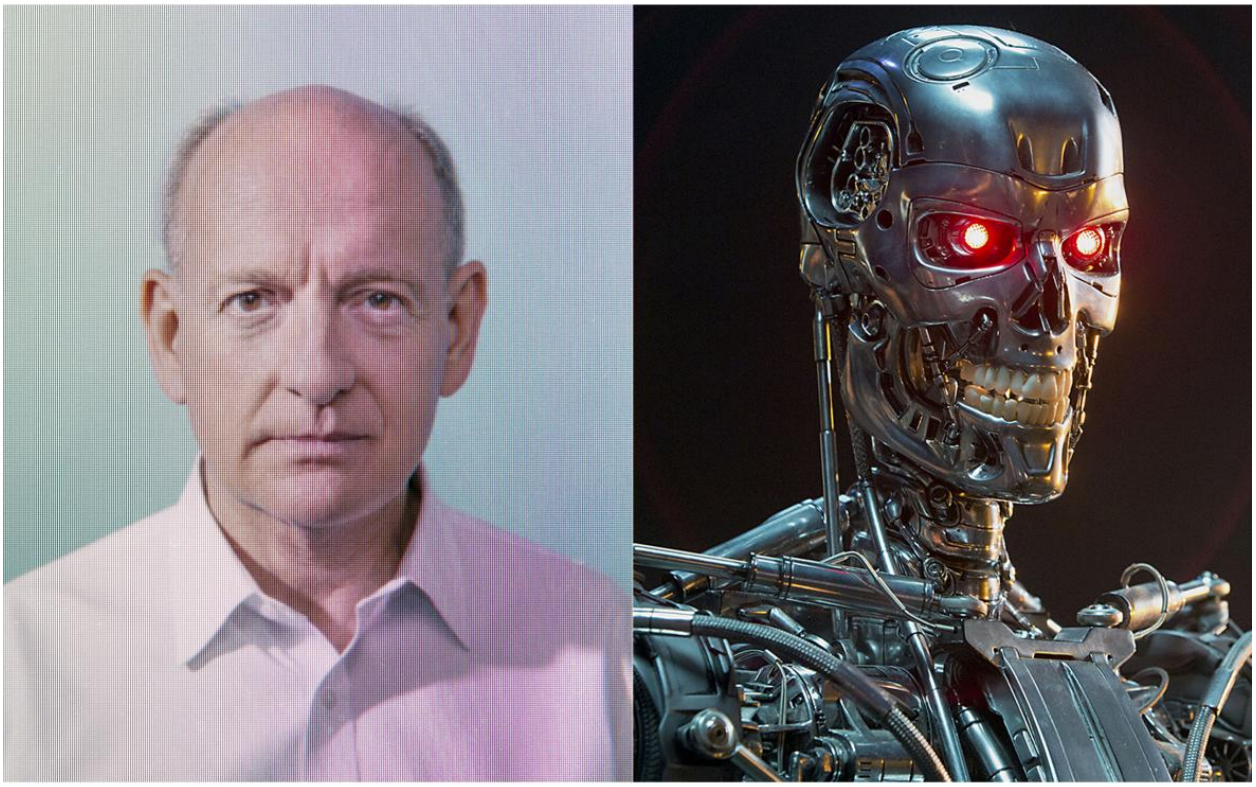
Ask that algorithm to do literally anything else, however, and it is utterly powerless. Such "tool AI", Russell says, "couldn't plan its way out of a paper bag". This is why the industry, at least outwardly, is rather blasé about the threat, or even the possibility, of general intelligence. A Google executive confided recently that for years, the search giant's auto-complete AI would turn every one of his emails to chief executive Sundar Pichai from "Dear Sundar" to "Dear Sugar". It made for some awkward conversations.

There are still many breakthroughs, Russell admits, that are needed to take AI beyond narrow jobs to create truly super-intelligent machines that can handle any task you throw at them. And the possibility that a technology so powerful would ever come to fruition just seems, well, bonkers.

Scott Phoenix, founder of the Silicon Valley AI start-up Vicarious, explains what it might look like when (if?) it arrives: “Imagine a person who has a photographic memory and has read every document that any human has ever written. They can think for 60,000 years for every second that passes. If you have a brain like that, questions that were previously out of reach for our small minds — about the nature of the universe, how to build a fusion reactor, how to build a teleporter — are suddenly in reach.”

Fantastical, you might think. But the same was once said of nuclear fission, Russell points out. The day after Lord Rutherford dismissed it as “moonshine” in 1933, another physicist, Leo Szilard, worked out how to do it. Twelve years later, Hiroshima and Nagasaki were levelled by atom bombs.

So, how long do we have before the age of superintelligent machines? Russell reckons they will arrive “in my kid’s lifetime”. In other words, the next 70 or 80 years.



He'll be back: Stuart Russell despairs at the overuse of Terminator imagery
DAVID VINTINER FOR THE SUNDAY TIMES MAGAZINE/REX

That is not an invitation to relax. For one, Russell admits, he is probably wrong. Trying to predict technological leaps is a mug's game. And neither will it be a "big bang" event, where one day we wake up and Hal 9000 is running the world. Rather, the rise of the machines will happen gradually, through a steady drumbeat of advances.

He points to the example of Yann LeCun, Facebook's chief AI scientist. In the 1990s, while he was at AT&T Labs, LeCun began developing a system to recognise handwriting. He cracked it. But only after he solved three other "holy grail problems" in the process: speech recognition, object recognition and machine translation. This is why, Russell argues, AI denialists are themselves in denial.

Few people may actually be working on general intelligence per se, but their siloed advances are all dropped into the same soup. "People talk about tool AI as if, 'Oh, it's

completely safe, there's nothing harmful about a Go program or something that recognises, you know, tumours in x-rays.' They say they don't have any connection with general-purpose AI. That's completely false," he says. "Google was founded to achieve human-level AI — the search engine is just how they get funds to finance the long-term goal."

Which is why we must start working — now — not just on how we overhaul AI, but society itself. We'll cover the former first.

The way algorithms work today — the way Russell has taught thousands of students to do it — is simple. Specify a clear, limited objective, and the machine figures out the optimal way to achieve it.

It turns out this is a very bad way to build AI. Consider social media. The content-selection algorithms at Facebook, YouTube, Twitter and the rest populate your feed with posts they think you'll find interesting, but their ultimate goal is something else entirely: revenue maximisation.

It turns out the best way to do that is to get you to click on advertisements, and the best way to do that is to disproportionately promote incendiary content that runs alongside them. "These simple machine-learning algorithms don't understand anything about human psychology, but they are super-powerful because they interact with you for hours and hours a day, they interact with billions of people and, because they have so much contact with you, they can manipulate you," Russell says. "They can manipulate your mind, your preferences, so that you are a different person, one who's more predictable." And it has worked a treat. The problem is that those algorithms do a lot more than stuff the pockets of Silicon Valley techies. They have also helped fuel "the resurgence of fascism, the dissolution of the social contract that

underpins democracies around the world, and, potentially, the end of the European Union and Nato", Russell writes. "Not bad for a few lines of code."

Our inability to see around every corner is what is wrong with AI today, Russell argues, but it's not a new problem. Humans have never been good at knowing what they actually want. He points to the story of King Midas, which goes back thousands of years. The mythical ruler got just what he wanted — that everything he touched turn to gold. He didn't bargain for the fact that this included his wine, his food, his family. With AI, it is no different. No matter how hard we try, we simply can't perfectly define objectives. There are always, as Donald Rumsfeld famously said, "unknown unknowns". And with something with potential global reach, there are no "do-overs" (second chances), Russell says.

Imagine, for example, that the era of general AI has arrived, and we ask it to do the heretofore impossible: to cure cancer. Huzzah! You might think this marks the start of a golden age of humanity. Not so fast, warns Russell.

"Within hours, the AI system has read the entire biomedical literature and hypothesised millions of potentially effective but previously untested chemical compounds," Russell writes. "Within weeks, it has induced multiple tumours of different kinds in every living human being so as to carry out medical trials of these compounds, this being the fastest way to find a cure. Oops."

How about we ask it to reverse the acidification of the oceans? Also not a top result. "The machine develops a new catalyst that facilitates an incredibly rapid chemical reaction between ocean and atmosphere and restores the oceans' pH levels. Unfortunately, a quarter of the oxygen in the atmosphere is used up in the process, leaving us to asphyxiate slowly and painfully. Oops."

You get the idea. But fear not. Russell has come up with a different way to build these tools. Instead of giving limited, specific objectives, the starting point would be much more vague: “Simply define the goal as ‘Be helpful to humans’,” he says. The path to doing so is, obviously, less clear, so the AI would be required to suss out how best to do that by constantly asking questions and observing human behaviour.

That subtle shift, Russell says, would mean there will be no such thing as killer AI because its whole reason for being would be to serve us. If it suddenly started killing us off, it would very happily pull the plug on itself. So the theory goes. Russell expects some pushback: “There will be some resistance because you’re kind of telling people, ‘OK, we think your foundations are wrong.’ My guess is that in 10 years’ time, people will say, ‘Well, of course we always thought this was just the way it is.’ ”

It sounds implausible, I argue. If AI is so vastly superior to us, can we really expect it to happily continue to work for us? Remember, we’re the gorillas in this scenario.

Russell demurs. Assuming that machines will act as we have towards lesser species is, apparently, a leap only a tiny human mind would make. “We have absolutely no idea what consciousness is, or how it functions in humans,” Russell says. “And no one is doing any research on how to make conscious machines, at least none that makes any sense to me. If you gave me a trillion dollars to build a conscious machine, I’d just give it back because I’d have absolutely no idea where to start.”

For someone who spends so much time future-scaping, one might have thought that Russell’s house would be suitably Jetson-like. It’s not. The walls are festooned with pastoral paintings. The rugs are a pale yellow, the sitting-room chairs stout and floral. If it were not so clean, it would feel almost fusty.

And it is frigid. By the end of the interview, I can't feel my toes. Russell, looking rather comfortable in his half-zip jumper, breezily drops bombshells about the future of humanity between sips of his ice-cold fizzy water.

I ponder my frozen phalanges. If I were a robot, trivialities such as needing to eat, or wearing thicker socks, would be of no concern. I could happily carry on for hours in subzero temperatures, and I'd probably ask better questions.

And that's the problem. Of all the scary things that AI heralds, an end to work as we know it is the most pressing and, among politicians, the most popular concern. The Democratic candidate Andrew Yang has based an entire US presidential campaign on his plan to deal with the new age of mass unemployment. The issue is also near the top of Boris Johnson's mind, though he has instead focused on "helpful robots washing and caring for an ageing population".

Most agree, however, that smart machines are quickening their bloodless march across not just blue-collar jobs but white-collar areas such as transport, law and medicine too. The accountancy firm PWC predicted recently that nearly a third of British jobs could be automated away within 15 years. Russell reckons that PWC's rather gloomy forecast may be underselling it. "If you just continue the status quo, the likely result is that most people would have no economic function," he says. Indeed, he argues, that process is well under way. A recent study by the Brookings Institution found that between 1980 and 2016 — the period that included the rise of the personal computer and then the internet — 54m net jobs were created in America. The problem? Most of those new jobs were not as good as the previous ones. They are not as well paid, require fewer skills and are less secure.

A hollowing out of so-called "middle-skilled" work, which requires a modicum of expertise but involves repetitive tasks, is gathering pace. "We have to engineer a

vision of a desirable future where machines are doing most of the work that we currently call work," Russell says.

Reboot camp: Elon Musk's Neuralink plans to implant chips into the human brain
REUTERS

There is no shortage of wild ideas that grapple with this future. Yang, the dark-horse presidential candidate, has proposed a "freedom dividend", his very American name for a $1,000 monthly stipend to every person over 18, to help cover the basics of life in a world where decent work is increasingly rare. Silicon Valley is obsessed with this idea, which is more widely known as universal basic income (UBI). Russell calls it something else: "an admission of failure".

Elon Musk goes several steps further, of course. His vision is to merge us with AI by implanting chips into our skulls, jacking us directly into the matrix. His company Neuralink has invented a brain drill that Musk reckons will allow us to "achieve a sort

of symbiosis with artificial intelligence". Russell doesn't like that idea either. "If everyone in the world has to have brain surgery to survive, then we made a mistake."

So, what is his plan to save the human race? Russell holds up his iPhone. "This represents a trillion dollars of research and development," he says. "How much have we put into how to make people happy? The fraction going to understanding the mind and happiness, into what makes a fulfilling life, which is after all what we really want, has been very small."

He has a point. In a world where work as we know it goes away, where creations far superior to us do all of life's heavy lifting, what does one do? From where does one derive satisfaction? Self-worth? Money?

Russell is calling for a new discipline: happiness engineering. "We have to learn to be better humans," he says. "People aren't going to have the wherewithal to really have a high-value occupation if we don't do that research, and if we don't then create the education systems around it — the training, the professions, the credentials. If we started now, it would take decades, and we aren't starting. So …" He trails off.

Just before we meet, Russell had been on a call, corralling a group of economists, AI researchers and science-fiction writers. The goal of that odd grouping was to come up with better ideas of how to cope with the world he is convinced is barrelling toward us.

"Economists are pretty pessimistic, but economics is not really a synthetic discipline, in the sense that it doesn't invent new economies on a regular basis. Whereas science fiction writers, that's kind of what they do," Russell says. "I'm hoping that by putting the groups together, the economists can bring realism and the writers can imagine ways that things could be different."

Things such as, say, elective brain surgery for all? Slaughterbots roaming the skies? A world where no one ever has to work? Sounds far-fetched, doesn't it? One thing is certain. Russell will not be happy with this newspaper. A recent review of his book was accompanied by a Terminator robot.

Oops.