The Telegraph

◆ Premium

› Techn... ›

Meet the man with the 'off-switch' for when the robots come for us

Professor Stuart Russell thinks he has worked out how to stop robots getting out of control

Follow

By **Harry de Quetteville**

1 OCTOBER 2019 • 6:00AM

Nearly 10 years ago Stuart Russell, one of the world's leading experts on artificial intelligence, was on the Paris Metro in Paris when he touched upon what he thinks may be the key to combating any potential threat posed to us by AI, while listening to Samuel Barber's *Agnus Dei*.

"It is an amazing piece," he says, in the soft but gently insistent fashion of a reserved man with immense authority in his field. "It made me think about what's important in the quality of human experience. I was having such an experience at the time. And it occurred to me that if artificial intelligence has a purpose it is to improve the overall quality of human experience.

"Then I thought: 'But AI doesn't know what the quality of human experience is.' And I realised that this was an interesting new way of thinking about AI. The idea came to me in a flash. So I tucked it away."

Depending on how the rise of superintelligent machines plays out - the head of the Armed Forces General Sir Nick Carter, for example, has just warned that AI weaponry is developing so quickly that in the future humans may have limited control on how we fight battles - Russell's epiphany may be celebrated as the moment that humanity began the practical process of defending itself from the most powerful adversary it has ever faced.

A slew of books over recent years - notably Max Tegmark's *Life 3.0* and Nick Bostrom's *Superintelligence* - have warned of the dangers of runaway AI escaping our control and - literally or metaphorically - turning us all into paperclips.

From social media algorithms to predictive text, AI has increasingly become part of our lives CREDIT: JAVIER PIERINI/GETTY

But it is Russell, in his new book *Human Compatible, AI and the Problem of Control*, who has begun to calculate some solutions to this looming crisis, which we must solve, not just to survive, but also to harness the immense possibilities for good of AI.

As Russell, who runs the Centre for Human-Compatible Artificial Intelligence (CHAI) at UC Berkeley in America, began mulling control mechanisms for overmighty machines of the future, he realised that the entire field of AI had made a logical, but profoundly dangerous, mistake.

"Humans," he writes, "are intelligent to the extent that our actions can be expected to achieve our objectives.'’

What went wrong was that machines became deemed intelligent by the same criteria - "to the extent that their actions can be expected to achieve their objectives".

But this way of thinking, Russell concluded, was a catastrophe because a) it gave up control over the means machines might use to achieve their ends, and b) humans are often uncertain or inept in setting out their desires in the first place.

The unintended effects, he writes, are already apparent. Social media algorithms given the simple, unmalicious objective of maximising clicks online do so not as we might imagine - by giving us stuff suited to our existing, vague, hard-to-predict preferences - but by giving us stuff at the fringes of our belief systems, constantly nudging us to the political extremes. They do so not because their creators want more extremism, but because the more predictable we are, the easier it is to provide links we will click on. And extremists, it turns out, are very predictable.

"The consequences include the resurgence of fascism," notes Russell. "Not bad for a few lines of code. Now imagine what a really intelligent algorithm would be able to do."

As a result he came up with an entirely new formulation to govern AI: "Machines are beneficial [bold] to the extent that their actions can be expected to achieve our [bold our] objectives."

It is a just a few words. But in it, he thinks, lies the logical, mathematical kernel that will keep man in charge of superintelligent machines.

He says: “My...principles are not laws for robots to follow. They are guarantees that the software is beneficial to you. We show it will allow itself to be switched off.

“If you can’t switch a machine off, the game is over.”

To find, as he delicately puts it, that “my own field of research posed a potential risk to my own species” was no great shock to Russell. “I read a lot of science-fiction as a child. Watched *2001*. Hal is a threat to the human astronauts. The theme of computers taking control was something I was aware of. But did I connect what I was doing to those questions? No.”

Artificial Intelligence | 4 Fictional Robots

1 Isaac Asimov’s *I, Robot*, a series of short stories written in 1950 that shaped our modern understanding of Artificial Intelligence (AI), is responsible for the ‘Three Laws of Robotics’, used by fiction writers and AI scientists alike. The most famous of the stories — which was turned into a film starring Will Smith in 2004 — is set in 2035, and sees an army of robots stage a revolution in a bid to enslave the human race. They ultimately fail.

2 Ridley Scott’s classic *Blade Runner* (1982) helped to forge our modern conception of the murderous robot. The film’s replicants look and sound just like humans – something some AI scientists say is unlikely – raising questions about whether they should be entitled to the same rights as people. Predictably, the robots turn evil and try to kill their creators. Denis Villeneuve’s sequel in 2017 was timely, coming amid a surge in AI paranoia.

3 Stanley Kubrick’s *2001: A Space Odyssey* (1968), inspired by an Arthur C. Clarke short story, introduced viewers to the emotionless, malevolent robot in the form of the HAL 9000 computer, voiced by Canadian actor Douglas Rain. HAL spies on the human astronauts by reading their lips, and sets one man adrift in space. The reasons for HAL’s malfunction have been a topic of debate among sci-fi aficionados ever since.

4 In *The Terminator* (1984), Arnold Schwarzenegger plays a cyborg assassin sent back in time from 2029 to kill a woman whose son will later become a saviour against the evil machines. He can drive motorcycles better than any human racer, and even mimics human voices. The Terminator ultimately fails, but that didn't prevent Hollywood from making five sequels.

That was largely because in the early Eighties, when Russell was completing his studies at Oxford, clunking technology made the very idea of machines approaching human-level intelligence seem like a joke.

Indeed, when Russell told his professors at Oxford that he had been accepted by Stanford to do a Phd in AI, they couldn't contain their amusement. "The very idea was literally laughable," he says.

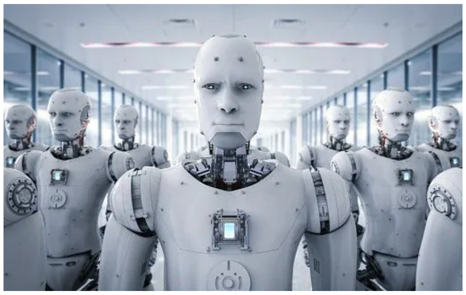
He did not bend before the sneers. A precocious talent, he was already well used to taking his scholastic destiny into his own hands. At his Lancashire prep school his maths teacher told him "you've done all the maths we have" and gave him textbooks to teach himself. Russell was 10. At St Paul's, aged 12, he rebelled from the Classics stream to focus on science and sought out a technology course (from Twickenham College) which the great public school did not offer. His interest in AI was alive when he was 15.

At Oxford, Russell expected to pursue an academic career as a particle physicist. "There are two questions," he says. "How does the universe work? And how does the mind work; how is intelligence possible?"

The key to the first, fundamental physics, was in a trough at the time. So he applied to Stanford and set about answering the second instead.

But that quickly led him to another question: What if we succeed in creating intelligent machines? "It was clear in the Eighties that no one in the field was thinking about that, which was a little disturbing to me."

Such complacency has now largely evaporated, partly due to books like Russell's and the widely publicised comments of those in a position to know, like Elon Musk and Stephen Hawking (who famously said AI "could spell the end of the human race").



Some AI experts have warned of the technology speeding up at a rate we can't keep up with CREDIT: PHONLAMAI

These days, Russell says, he is sometimes approached as if he is about to press the red nuclear button. "I've had people say 'Take your hands off the keyboard!' If you do any more AI research you're putting humanity in danger."

But to him, it is more, not less research that is needed. And the clock is ticking. Because AI advances in a very unpredictable way. A few years ago, if you tried giving vocal instructions to your phone, or getting it to translate something, the results would have been comic. Now the results are almost frighteningly good, because speech recognition and machine translation (along with visual object recognition), once deemed major AI challenges, have suddenly been mastered.

Russell thinks there are “at least half a dozen significant steps [still] to get over” before we create superintelligent machines. But if past breakthroughs are any guide, that could happen “overnight”.

“I know some serious, prominent researchers, who say we only have five years,” says Russell. Not long to cobble together safeguards which must be infallible, first time round.

His book sets out how that might be done through ensuring, say, that machines do not assume they know precisely what we want them to do, and to ask before acting. Baking in that uncertainty principle also provides a mathematical mechanism to guarantee the greatest safeguard of all: the off-switch. And all this because of the on-switch which piped Barber’s *Agnus Dei* into Russell’s headphones all those years ago back in Paris. One day we might all be grateful for that, altogether humbler, machine.

: Human Compatible by Stuart Russell (RRP £25). Buy now for £19.99 at books.telegraph.co.uk (https://books.telegraph.co.uk/ Product/Stuart-Russell/Human- Compatible--AI-and-the- Problem-of-Control/23803273) or call 0844 871 1514

Contact us

About us (https://corporate.telegraph.co.uk/)

Rewards

Archive (https://www.telegraph.co.uk/archive/)

Reader Prints (http://telegraph.newsprints.co.uk/)

Branded Content

Syndication and Commissioning

Guidelines

Privacy

Terms and Conditions

Advertising terms (http://spark.telegraph.co.uk/toolkit/advertising/terms-and-conditions/)

Fantasy Sport (https://fantasyfootball.telegraph.co.uk)

UK Voucher Codes (https://www.telegraph.co.uk/vouchercodes)

Modern Slavery