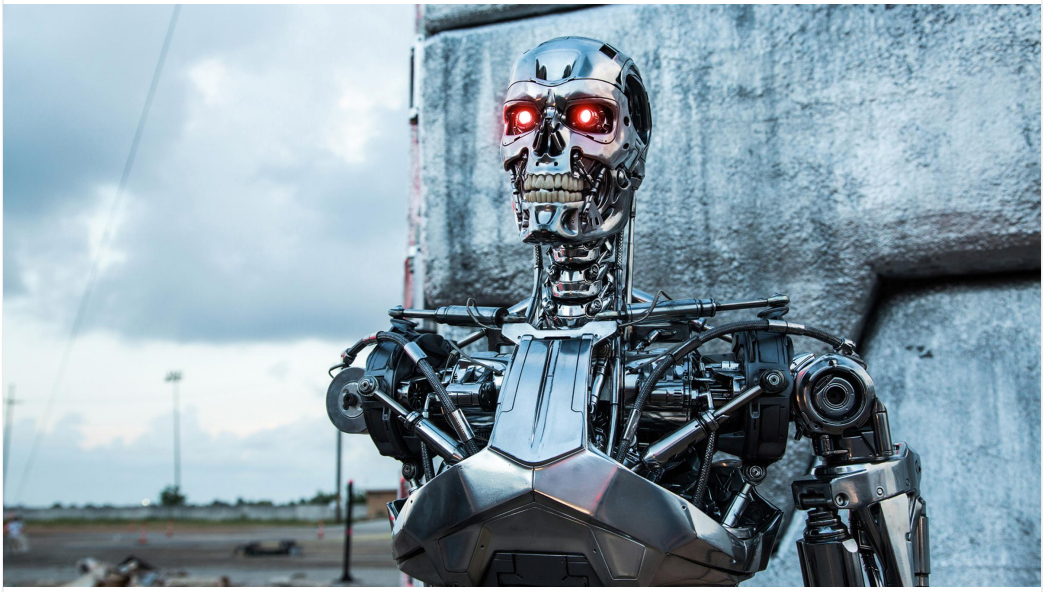
AI

Human Compatible by Stuart Russell review — an AI expert's chilling warning

A leading expert shows how serious the threat is from artificial intelligence. Review by James McConnachie



Killing machine A robot in Terminator Genisys (2015)
MELINDA SUE GORDON/PARAMOUNT PICTURES

The Sunday Times, October 6 2019, 12:01am

Every time Elon Musk or Bill Gates warns that humanity could soon be destroyed by rampaging hyperintelligent machines, I comfort myself with the thought that

I am not comforted any more. Stuart Russell calls the AI industry attitude “a kind of denialism” and argues that AI threatens our species. And he is professor of computer science at Berkeley, California, and one of the world’s leading experts in AI.

He invites us to imagine the UN receiving an email from an advanced alien civilisation. We are on our way, the email warns, and will land in 30 to 50 years. “Pandemonium”, he says, would hardly describe the reaction. But this is where things stand. We do not know when a superhumanly intelligent machine will be developed, but it will be. Probably within the lifetimes of our children.

Both Bill Gates, left, and Elon Musk have warned that humanity could soon be destroyed by AI ALAMY

Its behaviour might be unpredictable. It might use its abilities — digesting everything ever written in a single morning, say — to construct ever-more intelligent versions of itself. The resulting “intelligence explosion” might be rapid and unstoppable. You want to “pull the plug”? Naive, says Russell. If an AI is to

Russell does admit that several significant conceptual breakthroughs are still needed. This part of the book is difficult, but it is clear that researchers do not yet understand how reality can be grasped, and actions planned, at multiple levels of abstraction. For a robot even to work out how to walk, for instance, it first needs to "discover for itself that *standing up* is a thing". And it only gets more complicated.

Still, it would be a mistake to bet against human or AI ingenuity. Other scientific leaps, Russell observes, were also once thought to be impossible. Take nuclear fission. The morning after the nuclear physicist Lord Rutherford said that creating power from atomic energy was "moonshine", the physicist Leo Szilard dreamt up the chain reaction. That was in 1933. Within 12 years we had nuclear bombs.

Of course, just because nuclear fission is possible this does not mean superhuman AI will be. And here, Russell sidesteps the biggest problem: consciousness. In this area, he says, "we really do know nothing, so I'm going to say nothing", adding that "no behaviour has consciousness as a prerequisite". No behaviour, maybe, but if machines are to be purposeful, as well as merely

To be fair, the AIs that already trounce humans at Go or Chess are not consciously playing the game. Your self-driving car does not "know" your destination when it takes you there. Russell is not warning of the dangers of conscious machines, just that superintelligent ones might be misused or might misuse themselves.

AI is already dangerous. Israel has developed an autonomous "loitering munition" called Harop, which can hunt and destroy objects it classes as hostile. Antipersonnel microdrones equipped with facial-recognition systems and explosive weaponry might already exist. Slaughterbots, they are called.

On a more insidious level, smart speakers and mobile-phone assistants are "Trojan horses for AI", modelling our daily lives for the benefit of tech corporations. Social-media algorithms are already driving behavioural change, extremist content being more likely to generate the clicks the algorithm seeks to elicit. And rising intemperance and extremism are only accidental by-products; Russell asks us to "imagine what a *really* intelligent algorithm would be able to do".

So what can we do? Russell offers a fairly cursory survey of utilitarianism — which seems to be the only ethical show in town for computer scientists. And he insists on "three principles" of AI design reminiscent of Isaac Asimov's Three

uncertain about what those are so that it keeps asking. And it must seek to understand these preferences by observing human behaviour.

For all that to work, though, the leading tech firms in Silicon Valley and China must learn to accept regulation. “Let’s hope it doesn’t require a Chernobyl-sized disaster (or worse)”, he warns, “to overcome the industry’s resistance.”

This is not quite the popular book that AI urgently needs. Its technical parts are too difficult, its philosophical ones too easy. But it is fascinating, and significant. The AI industry is starting to accept that moving fast and breaking things is unwise if the thing that might break is humanity itself.

Human Compatible: AI and the Problem of Control by Stuart Russell
Allen Lane £25 pp352

Science | Middle East | Technology | United States

Related articles

SCIENCE

Novacene: The Coming Age of Hyperintelligence by James Lovelock review — an optimistic outlook from an unusual mind

Review by James McConnachie

Imagine that there are hyperintelligent beings who inhabit the Earth alongside us. They are...

June 30 2019

SOCIETY

Superior by Angela Saini review — the return of race

The most popular scientist in the world, the astrophysicist Neil deGrasse Tyson, runs the Hayden...

May 26 2019

FICTION

Machines Like Me by Ian McEwan — a robot love triangle

Johanna Thomas-Corr

At the age of 70, Ian McEwan is in a phase of late-career puckishness. His previous...

April 10 2019

Newest

S

Add to the conversation...

D

DNC 2 HOURS AGO

“no behaviour has consciousness as a prerequisite”.
yes, consciousness is just another software loop in our brains, and side issue.

Reply Recommend Report

BACK TO TOP

The Times Academy
Times Print Gallery
Times Crossword Club
Sunday Times Driving
Times+
The Sunday Times Rich List
Times Expert Traveller
Good University Guide
Schools Guide
Newsletters
Best Places to Live
Best Places to Stay
Announcements
Times Appointments
Podcasts

Privacy & cookie policy
Licensing
Cookie settings
Site map
Topics
Commissioning terms
Terms and conditions