QILAI SHEN/BLOOMBERG VIA GETTY

Surveillance cameras at the 2019 World Artificial Intelligence Conference in Shanghai, China.

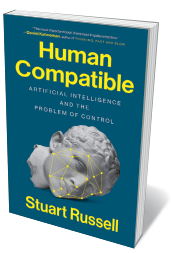
ARTIFICIAL INTELLIGENCE

Raging robots, hapless humans: the AI dystopia

Stuart Russell's latest probes the future of machine learning. **David Leslie** critiques it.

In *Human Compatible*, his new book on artificial intelligence (AI), Stuart Russell confronts full on what he calls "the problem of control". That is, the possibility that general-purpose AI will ultimately eclipse the intellectual capacities of its creators, to irreversible dystopian effect.

The control problem is not new. Novelist Samuel Butler's 1872 science-fiction classic *Erewhon*, for instance, features concerns about robotic superhuman intelligences that enslave their anthropoid architects, rendering them "affectionate machine-tickling aphids". But, by 1950, Norbert Wiener, the inventor of cybernetics, was writing (in *The Human Use of Human Beings*) that the danger to society "is not from the machine itself but from what man makes of it". Russell's book in effect hangs on this tension: whether the problem is controlling the creature, or the creator. In a sense, that has been at the core of AI from its inception.



Human Compatible: Artificial Intelligence and the Problem of Control
STUART RUSSELL
Viking (2019)

Even in its infancy, AI was swaddled in bitter controversy. Russell briefly touches on the moment its inventors convened at a 1956 workshop at Dartmouth College in Hanover, New Hampshire. Here, in the legendary birthplace of AI, they quarrelled over what to call their still-slumbering creation. Polymath and future Nobel laureate Herbert Simon, and computer scientist Allen Newell, favoured the name "complex information processing". The precision of the moniker evoked the restraint of the modern scientific method, harking back to the brick-by-brick processes of discovery exemplified by the likes of James Clerk Maxwell. Computer scientists John McCarthy and Marvin Minsky (let's call them the Intelligentists) favoured the muddier "artificial intelligence". For McCarthy, it had marketing value. For Minsky, defining it was "more of an aesthetic question or one of [a] sense of dignity, than a technical matter".

McCarthy, Minsky and other Intelligentists had bought in to behaviourism, a field straddling the natural and human sciences and offering access to a rich psychological vocabulary. They seemed to assume that the appearance of 'minded' behaviour was logically sufficient to prove its existence. Thus

they could claim that their machines could think and perceive simply because they looked as if they did.

As we know, Intelligentist nomenclature won out. That cast the die for a field that has faced recurrent patterns of overpromise and under-delivery, hubris and long, wound-licking winters of discontent. One wonders what the fate of AI research might have looked like had Simon and Newell's handle prevailed. Would Nick Bostrom's best-selling 2014 book *Superintelligence* have had as much play had it been called *Super Complex Information Processing Systems*? And would Russell have even written this book?

That point remains debatable. What is certain is that *Human Compatible* marks a major stride in AI studies, not least in its emphasis on ethics. At the book's heart, Russell incisively discusses the misuses of AI. He warns about how, deployed in combination with invasive data collection, AI applications such as voice and facial-recognition technologies, deepfake generators and information-integration systems can be used for surveillance, control and mass-behavioural manipulation. Stressing human vulnerability to such technologies, he emphasizes the right to the mental security of living "in a largely true information environment". And he makes a persuasive argument for rejecting lethal autonomous weapons as "scalable weapons of mass destruction".

It might seem surprising, in a solid, cautionary account of contemporary misuses and abuses of AI, that Russell fails to do justice to current boots-on-the-ground benefits. These are already helping people to tackle challenges such as climate change, the biodiversity drain, disease detection and disaster relief. For instance, AI applications in medical image analysis have advanced the early diagnosis of breast cancer, non-small-cell lung cancer and chronic heart failure. And machine-learning models that integrate macro-scale climate prediction with local observational data allow scientists to assist governments in climate adaptation and mitigation.

But "boring" yet beneficial "tool AI", as Russell calls it, is not the book's protagonist. That is "superintelligent" AI. This is where the control problem truly comes in. It is to do with the impossibility of AI management once it exceeds the general cognitive capabilities of humans. We're not there yet, he assures us, but when it does happen, we'd better be prepared.

Russell, however, fails to convince that we will ever see the arrival of a "second intelligent species". What he presents instead is a dizzyingly inconsistent account of "intelligence" that will leave careful readers scratching their heads. His definition of AI reduces this quality to instrumental rationality. Rational agents act intelligently, he tells us, to the degree that their actions aim to achieve their objectives, hence maximizing expected utility. This is likely to please hoary behavioural economists, with proclivities for formalization, and AI technologists squeaking reward functions onto whiteboards. But it is a blinkered characterization, and it leads Russell into absurdity when he applies it to what he calls "overly intelligent" AI.

TOMOHIRO OHSUMI/GETTY

An artificial-intelligence personal assistant being demonstrated at a trade show in Japan in 2018.

Russell's examples of human purpose gone awry in goal-directed superintelligent machines are bemusing. He offers scenarios such as a domestic robot that roasts the pet cat to feed a hungry child, an AI system that induces tumours in every human to quickly find an optimal cure for cancer, and a geoengineering robot that asphyxiates humanity to deacidify the oceans. One struggles to identify any intelligence here.

> ***"AI applications can be used for surveillance, control and mass-behavioural manipulation."***

Still, Russell all but admits that instrumental aptitude is not enough to account for the full gamut of intelligence capability. Sacrificing the coherence of his own definition, he hedges his bets. He offers a wish list of "conceptual breakthroughs" that are still needed for the conjuring of superintelligent AI. It includes gaining a capacity for common sense, a grasp of context and relevance, and an understanding of natural language.

This inventory is difficult to stomach. It merely rehearses more than 60 years of unanswered criticisms, intractable shortcomings and repeated failures that have largely derived from what cognitive scientist Zenon Pylyshyn in 1987 called "problems of holism in reasoning". That is, problems hinging on the failure of AI systems to equal the remarkable human capacity to put together a working but ever-provisional understanding of the world from the infinite number of possible meanings, properties and relationships that together constitute its moving frame.

There is no hard evidence in *Human Compatible* on how complex information-processing systems, superintelligent or otherwise, would begin to be able to do this. Nor does Russell clarify how this capacity lines up with the success-obsessed *Robo economicus* species of superintelligence that he invokes. Instead, he ignores the strain of twentieth-century thinking whose holistic, contextual understanding of reasoning has led to a humble acknowledgement of the existential limitations of intelligence itself. As a consequence, Russell ultimately falls prey to the techno-solutionist idea that intelligence can be treated as an 'engineering problem', rather than a constraining dimension of the human condition that demands continuous, critical self-reflection.

Russell's book does, however, sound an important alarm bell, although it is not quite the one he intended. The real control problem isn't managing the coming of transcendent superintelligent creatures. More critically, it has to do with reining in the triumphalist creators who may be developing increasingly "autonomous" AI technologies under the auspices of the misguided definition of intelligence that Russell himself champions. ■

David Leslie *is the Ethics Fellow at The Alan Turing Institute in London.*
e-mail: dleslie@turing.ac.uk

CLARIFICATION

Owing to a typographical error, the book review 'Raging robots, hapless humans: the AI dystopia' (*Nature* **574**, 32–33; 2019), included the phrase "maximizing being demonstrated utility". It should have said "maximizing expected utility".