'RoboBees' are meant for artificial pollination but could have unforeseen environmental effects.

ARTIFICIAL INTELLIGENCE

The future is superintelligent

Stuart Russell weighs up a book on the risks and rewards of the AI revolution.

Max Tegmark is a renowned physicist. He is also the irrepressibly optimistic co-founder of the Future of Life Institute in Cambridge, Massachusetts (motto: "Technology is giving life the potential to flourish like never before ... or to self-destruct. Let's make a difference!"). Now, in *Life 3.0*, he tackles a pressing future development — the evolution of artificial intelligence (AI). He argues that the risks demand serious thought if our "cosmic endowment" is not to be inadvertently thrown away.

In the interests of disclosure, Tegmark and I are collaborators and share a literary agent. With physicists Stephen Hawking and Frank Wilczek, we wrote the 2014 *Huffington Post* article 'Transcending complacency on superintelligent machines' (see go.nature.com/2wadkao). Ostensibly a review of Wally Pfister's dystopian AI film *Transcendence*, this was really a call to the AI community to take the risks of intelligent systems seriously. Thus, I am unlikely to disagree strongly with the premise of *Life 3.0*. Life, Tegmark argues, may or may not spread through the Universe and "flourish for billions or trillions of years" because of decisions we make now — a possibility both seductive and overwhelming.

The book's title refers to a third phase in

evolutionary history. For almost 4 billion years, both hardware (bodies) and software (capacity for generating behaviour) were fixed by biology. For the next 100,000 years, learning and culture enabled humans to adapt and control their own software. In the imminent third phase, both software and hardware can be redesigned. This may sound like transhumanism — the movement to re-engineer body and brain — but Tegmark's focus is on AI, which supplements mental capabilities with external devices.

Tegmark considers both risks and benefits. Near-term risks include an arms race in autonomous weapons and dramatic reductions in employment. The AI community is practically unanimous in condemning the creation of machines that can choose to kill humans, but the issue of work has sparked debate. Many predict an economic boon — AI inspiring new jobs to replace old, as with

Life 3.0: Being Human in the Age of Artificial Intelligence
MAX TEGMARK
Knopf: 2017.

previous industrial revolutions. Tegmark wryly imagines two horses discussing the rise of the internal combustion engine in 1900. One predicts "new jobs for horses ... That's what's always happened before, like with the invention of the wheel and the plow." For most horses, alas, the "new job" was to be pet food. Tegmark's analysis is compelling, and shared by economists such as Paul Krugman. But the question remains: what desirable economy might we aim for, when most of what we now call work is done by machines?

The longer-term risks are existential. The book's fictional prelude describes a reasonably plausible scenario in which superintelligent AI might emerge. Later, Tegmark ranges over global outcomes from near-Utopias to human enslavement or extinction. That we have no idea how to steer towards the better futures points to a dearth of serious thinking on why making AI better might be a bad thing.

Computer pioneer Alan Turing, raising the possibility in 1951 that our species would at best be "greatly humbled" by AI, expressed the general unease of making something smarter than oneself. Assuaging this unease by curtailing progress on AI may be neither feasible nor preferable. The most interesting part of *Life 3.0* explains that the real issue

THIERRY FAUSE/LIGHTROCKET VIA GETTY

is the potential for misaligned objectives. Cybernetics founder Norbert Wiener wrote in 1960, “We had better be quite sure that the purpose put into the machine is the purpose which we really desire.” Or, as Tegmark has it, “It’s unclear how to imbue a superintelligent AI with an ultimate goal that neither is undefined nor leads to the elimination of humanity.” In my view, this technological and philosophical problem demands all the intellectual resources we can bring to bear.

Only if we solve it can we reap the benefits. Among these is expansion across the Universe, perhaps powered by such exotic technologies as Dyson spheres (which would capture the energy of a star), accelerators built around black holes or Tegmark’s theorized sphalerizers (like diesel engines, but quark-powered and one billion times more efficient). For sheer science fun, it’s hard to beat the explanations of how much upside the Universe and the laws of physics will allow. We may one day, for example, expand the biosphere “by about 32 orders of magnitude”. It’s seriously disappointing, then, to learn that cosmic expansion may limit us to settling only 10 billion galaxies. And we feel our descendants’ anxiety as “the threat of dark energy tearing cosmic civilizations apart motivates massive cosmic engineering projects”.

The book concludes with the Future of Life Institute’s role in moving these issues into mainstream AI thinking — for which Tegmark deserves huge credit. He is not alone, of course, in raising the alarm. In its sweeping vision, *Life 3.0* has most in common with Nick Bostrom’s 2014 *Superintelligence* (Oxford University Press). Unlike Bostrom, however, Tegmark is not trying to prove that risk is unavoidable; and he eschews dense philosophy in favour of asking the reader which scenarios they think more probable or desirable.

Although I strongly recommend both books, I suspect that Tegmark’s is less likely to provoke in AI researchers a common allergic reaction — a retreat into defensive arguments for paying no attention. Here’s a typical one: we don’t worry about remote but species-ending possibilities such as black holes materializing in near-Earth orbit, so why worry about superintelligent AI? Answer: if physicists were working to make such black holes, wouldn’t we ask them if it was safe?

The Economist has drily characterized the overarching issue thus: “The implications of introducing a second intelligent species onto Earth are far-reaching enough to deserve hard thinking.” *Life 3.0* is far from the last word on AI and the future, but it provides a fascinating glimpse of the hard thinking required. ■

Stuart Russell is professor of computer science at the University of California, Berkeley and co-author of *Artificial Intelligence: A Modern Approach*. e-mail: russell@berkeley.edu

Books in brief

Willy Ley: Prophet of the Space Age

Jared S. Buss UNIVERSITY PRESS OF FLORIDA (2017)

Among the scientists who fled Europe for the United States in the 1930s and 1940s was Willy Ley, first historian of spaceflight. In Jared Buss’s nuanced biography, the German-born rocket expert emerges as a spirited science educator whose promotion of space exploration paved the way for NASA’s triumphs. By 1960, Ley’s romanticized vision of science had been occluded by the harder-edged approach of his fellow émigré and collaborator, aerospace wizard Wernher von Braun — but only partly. As Buss reveals, Ley’s “reenchantment” of big science with a sense of wonder has held.

Uneasy Street: The Anxieties of Affluence

Rachel Sherman PRINCETON UNIVERSITY PRESS (2017)

There have been many cogent analyses of US income inequality. Sociologist Rachel Sherman’s welcome addition probes the psychology and socio-economics of affluence. Sherman interviewed 50 New Yorkers, from academics to financiers, with incomes between US\$250,000 and more than \$10 million. She found complex adaptations to privilege, from ethical acrobatics normalizing extreme consumption to defensive assertions that wealth is “earned”. Most of her interviewees focus on crafting a self-legitimizing moral universe rather than addressing the structural inequalities that trigger their unease.

What It's Like to be a Dog

Gregory Berns BASIC (2017)

The hyper-curiosity of a pet terrier spurred neuroscientist Gregory Berns to probe the dog’s brain — an experiment demanding that she stay awake and unrestrained in a magnetic resonance imaging scanner. Many more dogs and trials later, the project has generated numerous findings on canine brain function, such as how owners’ smells alone activate the reward response. This is just one “adventure in animal neuroscience” that Berns pursues, as he investigates beat synchronization in sea lions, auditory pathways in dolphins and the putative behaviour of the extinct thylacine.

Frankenstein Dreams

Michael Sims BLOOMSBURY (2017)

Victorian upheavals in science and technology jolted the era, and the imaginations of its speculative-fiction writers. This nightshade nosegay of vintage sci-fi tales and excerpts, plucked by writer Michael Sims, richly reflects the shifts. Along with miniature masterworks by the likes of Mary Shelley, Edgar Allan Poe, H. G. Wells and Ambrose Bierce are unexpected gems. William Henry Rhodes’s ‘The Telescopic Eye’, for instance, features chariot-wheel-shaped aliens called Lunarians; and Mary Wilkins Freeman’s ‘The Hall Bedroom’ is a hallucinatory journey into the fifth dimension.

DNA: The Story of the Genetic Revolution

James D. Watson, with Andrew Berry and Kevin Davies KNOPF (2017)

In this update of *DNA: The Secret of Life* (Knopf, 2003), James Watson — the controversially outspoken co-discoverer of DNA’s structure — remaps the genetic landscape. There are new chapters on personal genetics and cancer treatment; key discoveries, such as the CRISPR revolution kick-started by Emmanuelle Charpentier, Jennifer Doudna and Feng Zhang; and findings in epigenetics and agricultural chemistry. Backing nature over nurture, Watson ends by arguing (with caveats) in favour of future germline gene editing. **Barbara Kiser**