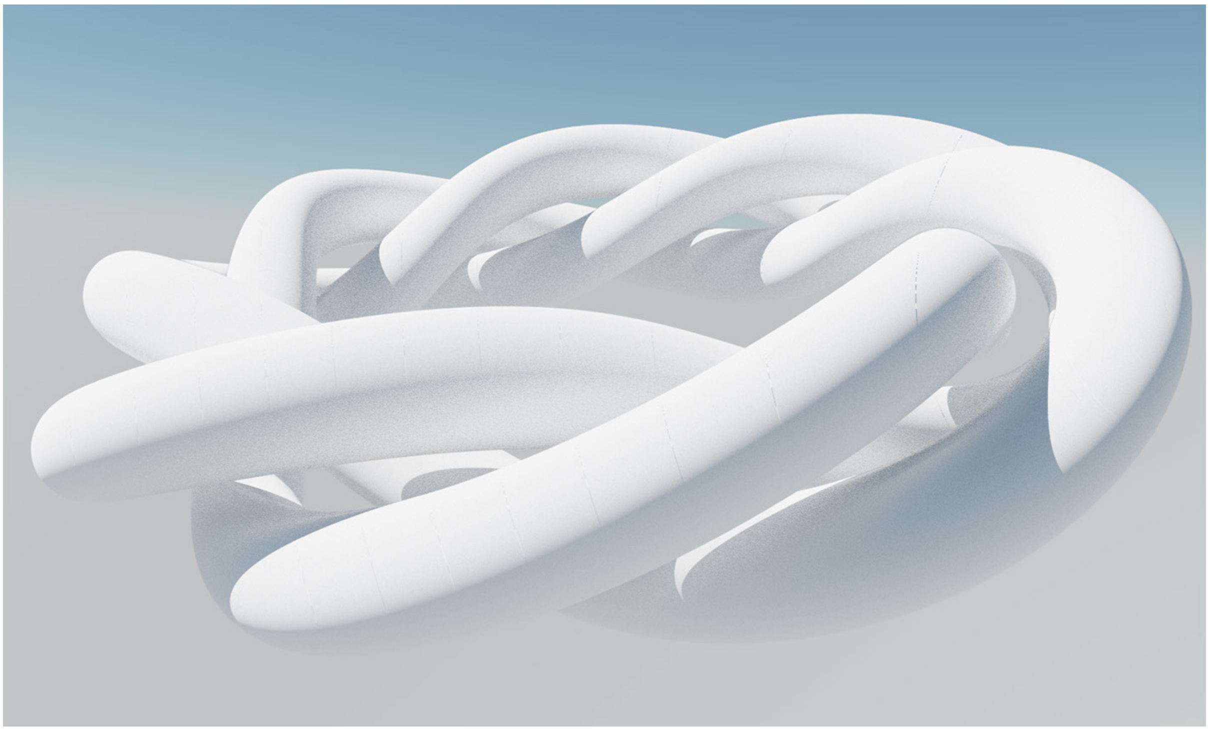
aeon

Now it's time to prepare for the Machinocene

Huw Price



Human-level intelligence is familiar in biological hardware – you're using it now. Science and technology seem to be converging, from several directions, on the possibility of similar intelligence in non-biological systems. It is difficult to predict when this might happen, but most artificial intelligence (AI) specialists estimate <http://philpapers.org/rec/MLLFPI> that it is more likely than not within this century.

Freed of biological constraints, such as a brain that needs to fit through a human birth canal (and that runs on the power of a mere 20W lightbulb), non-biological machines might be much more intelligent than we are. What would this mean for us? The leading AI researcher Stuart Russell suggests that, for better or worse, it would be 'the biggest event in human history'. Indeed, our choices in this century might have long-term consequences not only for our own planet, but for the galaxy at large, as the British Astronomer Royal Martin Rees has observed <http://www.newstatesman.com/sci-tech/2014/11/martin-rees-world-2050-and-beyond> . The future of intelligence in the cosmos might depend on what we do right now, down here on Earth.

Should we be concerned? People have been speculating about machine intelligence for generations – so what's new?

Well, two big things have changed in recent decades. First, there's been a lot of real progress – theoretical, practical and technological – in understanding the mechanisms of intelligence, biological as well as non-biological. Second, AI has now reached a point where it's immensely useful for many tasks. So it has huge commercial value, and this is driving huge investment – a process that seems bound to continue, and probably accelerate.

One way or another, then, we are going to be sharing the planet with a lot of non-biological intelligence. Whatever it brings, we humans face this future together. We have an obvious common interest in getting it right. And we need to nail it the first time round. Barring some calamity that ends our technological civilisation without entirely finishing us off, we're not going to be coming this way again.

There have been encouraging signs of a growing awareness of these issues. Many thousands of AI researchers and others have now signed an open letter <http://futureoflife.org/ai-open-letter/> calling for research to ensure that AI is safe and beneficial. Most recently, there is a welcome new Partnership on AI <https://www.partnershiponai.org/> to Benefit People and Society by Google, Amazon, Facebook, IBM and Microsoft.

For the moment, much of the focus is on safety, and on the relatively short-term benefits and impacts of AI (on jobs, for example). But as important as these questions are, they are not the only things we should be thinking about. I'll borrow an example from Jaan Tallinn, a founding engineer at Skype. Imagine that we were taking humanity into space in a fleet of giant ships. We would need to be sure that these ships were safe and controllable, and that everybody was properly housed and fed. These things would be crucial, but they wouldn't be enough by themselves. We'd also do our best to figure out where the fleet was going to take us, and what we could do to steer our way towards the best options. There could be paradise worlds out there, but there's a lot of cold, dark space in between. We'd need to know where we were going.

In the case of the long-term future of AI, there are reasons to be optimistic. It might help us to solve many of the practical problems that defeat our own limited brains. But when it comes to what the cartography of possible futures looks like, which parts of it are better or worse, and how we steer towards the best outcomes – on those matters we are still largely ignorant. We have some sense of regions we need to avoid, but much of the map remains *terra incognita*. It would be a peculiarly insouciant optimist who thought we should just wait and see.

One of the far-sighted writers who saw this coming was the great Alan Turing. '[I]t seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers,' he wrote at the conclusion of a 1951 lecture <http://uberty.org/wp-content/uploads/2015/02/intelligent-machinery-a-heretical-theory.pdf> . In his 1950 paper <http://mind.oxfordjournals.org/content/LIX/236/433.full.pdf+html> on the so-called Turing Test, designed to gauge our readiness to ascribe human-like intelligence to a machine, Turing closes with these words: 'We can only see a short distance ahead, but we can see plenty there that needs to be done.' We're well beyond Turing's horizon, but this progress does nothing to alleviate the sense that there are still pressing questions we must try to answer. On the contrary – we live among pressures that will soon take us beyond our own present horizon, and we have even more reason than Turing to think that what lies ahead could be very big indeed.

If we are to develop machines that think, ensuring that they are safe and beneficial is one of the great intellectual and practical challenges of this century. And we must face it together – the issue is far too large and crucial to be tackled by any individual institution, corporation or nation. Our grandchildren, or their grandchildren, are likely to be living in a different era, perhaps more Machinocene than Anthropocene. Our task is to make the best of this epochal transition, for them and the generations to follow. We need the best of human intelligence to make the best of artificial intelligence.

Stuart Russell and Martin Rees are affiliated with the new Leverhulme Centre for the Future of Intelligence at the University of Cambridge, where Huw Price is the academic director.

Martin Rees and Jaan Tallinn are co-founders of the Centre for the Study of Existential Risk at the University of Cambridge, where Huw Price is the academic director.

Huw Price is Bertrand Russell Professor of Philosophy and a fellow of Trinity College at the University of Cambridge. He is also academic director of the Centre for the Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence. His most recent book is Expressivism, Pragmatism and Representationalism (2013).

aeon.co 17 October, 2016