Are We Taking Artificial Intelligence Seriously Enough?

Stephen Hawking, Stuart Russell, Max Tegmark, Frank Wilczek
May 1, 2014
Independent (independent.co.uk)

Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks, says a group of leading scientists

With the Hollywood blockbuster *Transcendence* playing in cinemas, with Johnny Depp and Morgan Freeman showcasing clashing visions for the future of humanity, it's tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake in history.

Artificial-intelligence (AI) research is now progressing rapidly. Recent landmarks such as self-driving cars, a computer winning at *Jeopardy!* and the digital personal assistants Siri, Google Now and Cortana are merely symptoms of an IT arms race fuelled by unprecedented investments and building on an increasingly mature theoretical foundation. Such achievements will probably pale against what the coming decades will bring.

The potential benefits are huge; everything that civilisation has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools that AI may provide, but the eradication of war, disease, and poverty would be high on anyone's list. Success in creating AI would be the biggest event in human history.

Unfortunately, it might also be the last, unless we learn how to avoid the risks. In the near term, world militaries are considering autonomous-weapon systems that can choose and eliminate targets; the UN and Human Rights Watch have advocated a treaty banning such weapons. In the medium term, as emphasised by Erik Brynjolfsson and Andrew McAfee in *The Second Machine Age*, AI may transform our economy to bring both great wealth and great dislocation.

Looking further ahead, there are no fundamental limits to what can be achieved: there is no physical law precluding particles from being organised in ways that perform even more advanced computations than the arrangements of particles in human brains. An explosive transition is possible, although it might play out differently from in the movie: as Irving Good realised in 1965, machines with superhuman intelligence could repeatedly improve their design even further, triggering what Vernor Vinge called a "singularity" and Johnny Depp's movie character calls "transcendence".

One can imagine such technology outsmarting financial markets, out-inventing human researchers, out- manipulating human leaders, and developing weapons we cannot even understand. Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all.

So, facing possible futures of incalculable benefits and risks, the experts are surely doing everything possible to ensure the best outcome, right? Wrong. If a superior alien civilisation sent us a message saying, "We'll arrive in a few decades," would we just reply, "OK, call us when you get here – we'll leave the lights on"? Probably not – but this is more or less what is happening with AI. Although we are facing potentially the best or worst thing to happen to humanity in history, little serious research is devoted to these issues outside non-profit institutes such as the Cambridge Centre for the Study of Existential Risk, the Future of Humanity Institute, the Machine Intelligence Research Institute, and the Future of Life Institute. All of us should ask ourselves what we can do now to improve the chances of reaping the benefits and avoiding the risks.

Stephen Hawking is the director of research at the Department of Applied Mathematics and Theoretical Physics at Cambridge and a 2012 Fundamental Physics Prize laureate for his work on quantum gravity. Stuart Russell is a computer-science professor at the University of California, Berkeley and a co-author of 'Artificial Intelligence: A Modern Approach'. Max Tegmark is a physics professor at the Massachusetts Institute of Technology (MIT) and the author of 'Our Mathematical Universe'. Frank Wilczek is a physics professor at the MIT and a 2004 Nobel laureate for his work on the strong nuclear force.