Fear artificial stupidity, not artificial intelligence

Stephen Hawking thinks computers may surpass human intelligence and take over the world. We won't ever be silicon slaves, insists an AI expert

It is not often that you are obliged to proclaim a much-loved genius wrong, but in his alarming prediction on artificial intelligence and the future of humankind, I believe Stephen Hawking has erred. To be precise, and in keeping with physics – in an echo of Schrödinger's cat – he is simultaneously wrong and right.
Asked how far engineers had come towards creating artificial intelligence, Hawking replied: "Once humans develop artificial intelligence it would take off on its own and redesign itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded."

In my view, he is wrong because there are strong grounds for believing that computers will never replicate all human cognitive faculties. He is right because even such emasculated machines may still pose a threat to humankind's future – as autonomous weapons, for instance.

Such predictions are not new; my former boss at the University of Reading, professor of cybernetics Kevin Warwick, raised this issue in his 1997 book *March of the Machines*. He observed that robots with the brain power of an insect had already been created. Soon, he predicted, there would be robots with the brain power of a cat, quickly followed by machines as intelligent as humans, which would usurp and subjugate us.

Triple trouble

This is based on the ideology that all aspects of human mentality will eventually be realised by a program running on a suitable computer – a so-called strong AI. Of course, if this is possible, a runaway effect would eventually be triggered by accelerating technological progress – caused by using AI systems to design ever more sophisticated AIs and Moore's law, which states that raw computational power doubles every two years.

I did not agree then, and do not now.

I believe three fundamental problems explain why computational AI has historically failed to replicate human mentality in all its raw and electro-chemical glory, and will continue to fail.

First, computers lack genuine understanding. The Chinese Room Argument is a famous thought experiment by US philosopher John Searle that shows how a computer program can appear to understand Chinese stories (by responding to questions about them appropriately) without genuinely understanding anything of the interaction.

Second, computers lack consciousness. An argument can be made, one I call Dancing with Pixies, that if a robot experiences a conscious sensation as it interacts with the world, then an infinitude of consciousnesses must be everywhere: in the cup of tea I am drinking, in the seat that I am sitting on. If we reject this wider state of affairs – known as panpsychism – we must reject machine consciousness.
Lastly, computers lack mathematical insight. In his book *The Emperor's New Mind*, Oxford mathematical physicistRoger Penrose argued that the way mathematicians provide many of the "unassailable demonstrations" to verify their mathematical assertions is fundamentally non-algorithmic and non-computational.

Not OK computer

Taken together, these three arguments fatally undermine the notion that the human mind can be completely realised by mere computations. If correct, they imply that some broader aspects of human mentality will always elude future AI systems.

Rather than talking up Hollywood visions of robot overlords, it would be better to focus on the all too real concerns surrounding a growing application of existing AI – autonomous weapons systems.

In my role as an AI expert on the International Committee for Robot Arms Control, I am particularly concerned by the potential deployment of robotic weapons systems that can militarily engage without human intervention. This is precisely because current AI is not akin to human intelligence, and poorly designed autonomous systems have the potential to rapidly escalate dangerous situations to catastrophic conclusions when pitted against each other. Such systems can exhibit genuine artificial stupidity.
It is possible to agree that AI may pose an existential threat to humanity, but without ever having to imagine that it will become more intelligent than us.