HU3101: History and Philosophy of Science

Satvik Saha, 19MS154 November 26, 2021

Question 1. Artificial Intelligence (AI) is undoubtedly one of the great achievements of science. Yet, shortly before his death, Stephen Hawking said that within a hundred years or so, AI-powered robots will rule the earth, and humans will have to work as their slaves. He said the only way for humans to escape this dire consequence is to colonize Mars, for which the required technology is already available. Do you think Hawking was being unduly pessimistic? Or was he right?

Answer. In light of the breakneck speed of modern AI research, Stephen Hawking and other experts are right in asking 'when', rather than 'if' we will create a truly general purpose, powerful, and potentially malicious Artificial Intelligence.

Assuming that technology will indeed progress to this level, the possibility of this undesirable AI is enormous. One scenario is that AI is harmful by design – we have seen that science has been purposefully harnessed in service of atrocities like eugenics as well as chemical and nuclear warfare. The 'weaponization of AI' is a very real, serious concern, especially since cyber warfare, surveillance, and invasion of privacy by both tech companies and governments are commonplace today. A related idea is that AI may bring about the downfall of humanity by over-enthusiastically trying to fulfil our commands, only with unintended consequences. After all, the greatest strength and simultaneously the greatest weakness of a computer is that it does exactly what you tell it to do. If one tells an all-powerful AI to do something as benign as 'maximize human happiness', this gives it the incentive to put every human in a permanent sleep, delivering electrochemical signals of happiness directly to their brains (something like in the film *The Matrix*; whether this is particularly desirable is up to further debate!).

The far more likely scenario is an uncontrollable AI with goals that conflict ours. Hawking gives the analogy that when humans build over anthills, we do not act out of hate for ants; rather, they are simply beneath our thought[1]. On the other side, the ants may not even be aware of the goals, desires, or even existence of humans. Another thing to consider is that computers work much faster than humans; we use computer programs to build other computer programs, and so on. There will likely come a point where AI, in a quest to improve itself, gives birth to an even more powerful AI, and so on ad infinitum (much faster than biological evolution). This gives strength to the existence of arbitrarily powerful AI in the future.

In January 2015, Hawking and other experts signed *Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter*. It seems that active research into areas like machine ethics, algorithmic bias, and AI control (to name a few) are well worth it, indeed critical.

Hawking's 100 year deadline for colonizing Mars is in context of multiple doomsday scenarios, including climate change, disease, asteroid impacts, nuclear war, as well as an AI takeover. Making humanity an interplanetary civilisation would help reduce the chances of total extinction. Although we do *not* have all of the necessary technology to do this today (the effects of long term space travel along with radiation exposure on the human body is not properly

[1] The same argument has been made against initiating contact with intelligent alien life. If any alien civilisation has the technology to contact us, they are likely *much* more advanced than us, putting us in the position of the ants.

understood), we are rapidly heading in this direction. The goal of bringing down the cost of space travel and making it more accessible has received a lot of attention in particular, thanks to the involvement of private agencies such as SpaceX and Blue Origin.

In conclusion, Hawking's fears are well justified, and shared by many working on this problem. All solutions involve coordination and cooperation of researchers, policy makers, and the public on a global scale.