# Should we be afraid of AI?

## Machines seem to be getting smarter and smarter and much better at human jobs, yet true AI is utterly implausible. Why?

Suppose you enter a dark room in an unknown building. You might panic about monsters that could be lurking in the dark. Or you could just turn on the light, to avoid bumping into furniture. The dark room is the future of artificial intelligence (AI). Unfortunately, many people believe that, as we step into the room, we might run into some evil, ultra-intelligent machines. This is an old fear. It dates to the 1960s, when Irving John Good, a British mathematician who worked as a cryptologist at Bletchley Park with Alan Turing, made the following observation:

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.

Once ultraintelligent machines become a reality, they might not be docile at all but behave like Terminator: enslave humanity as a sub-species, ignore its rights, and pursue their own ends, regardless of the effects on human lives.

If this sounds incredible, you might wish to reconsider. Fast-forward half a century to now, and the amazing developments in our digital technologies have led many people to believe that Good's 'intelligence explosion' is a serious risk, and the end of our species might be near, if we're not careful. This is Stephen Hawking in 2014:

The development of full artificial intelligence could spell the end of the human race.

Last year, Bill Gates was of the same view:

I am in the camp that is concerned about superintelligence. First the machines will do a lot of jobs for us and not be superintelligent. That should be positive if we manage it well. A few decades after that, though, the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this, and don't understand why some people are not concerned.

And what had Musk, Tesla's CEO, said?

We should be very careful about artificial intelligence. If I were to guess what our biggest existential threat is, it's probably that… Increasingly, scientists think there should be some regulatory oversight maybe at the national and international level, just to make sure that we don't do something very foolish. With artificial intelligence, we are summoning the demon. In all those stories where there's the guy with the pentagram and the holy water, it's like, yeah, he's sure he can control the demon. Didn't work out.

The reality is more trivial. This March, Microsoft introduced Tay – an AI-based chat robot – to Twitter. They had to remove it only 16 hours later. It was supposed to become increasingly smarter as it interacted with humans. Instead, it quickly became an evil Hitler-loving, Holocaust-denying, incestual-sex-promoting, 'Bush did 9/11'-proclaiming chatterbox. Why? Because it worked no better than kitchen paper, absorbing and being shaped by the nasty messages sent to it. Microsoft apologised.

This is the state of AI today. After so much talking about the risks of ultraintelligent machines, it is time to turn on the light, stop worrying about sci-fi scenarios, and start focusing on AI's actual challenges, in order to avoid making painful and costly mistakes in the design and use of our smart technologies.

Let me be more specific. Philosophy doesn't do nuances well. It might fancy itself a model of precision and finely honed distinctions, but what it really loves are polarisations and dichotomies. Internalism or externalism, foundationalism or coherentism, trolley left or right, zombies or not zombies, observer-relative or observer-independent, possible or impossible worlds, grounded or ungrounded … Philosophy might preach the inclusive *vel* ('girls *or* boys may play') but too often indulges in the exclusive *aut aut* ('*either* you like it *or* you don't').

The current debate about AI is a case in point. Here, the dichotomy is between those who believe in *true* AI and those who do not. Yes, the real thing, not Siri in your iPhone, Roomba

# Should we be afraid of AI?

## Machines seem to be getting smarter and smarter and much better at human jobs, yet true AI is utterly implausible. Why?

in your living room, or Nest in your kitchen (I am the happy owner of all three). Think instead of the false Maria in *Metropolis* (1927); Hal 9000 in *2001: A Space Odyssey* (1968), on which Good was one of the consultants; C3PO in *Star Wars* (1977); Rachael in *Blade Runner* (1982); Data in *Star Trek: The Next Generation )(*1987); Agent Smith in *The Matrix* (1999) or the disembodied Samantha in *Her* (2013). You've got the picture. Believers in true AI and in Good's 'intelligence explosion' belong to the Church of Singularitarians. For lack of a better term, I shall refer to the disbelievers as members of the Church of AItheists. Let's have a look at both faiths and see why both are mistaken. And meanwhile, remember: good philosophy is almost always in the boring middle.