Can we stop AI outsmarting humanity?

The spectre of superintelligent machines doing us harm is not just science fiction, technologists say – so how can we ensure AI remains 'friendly' to its makers?

By Mara Hvistendahl Thu 28 Mar 2019

Photograph: Science Picture Co/Getty

(…) The term artificial intelligence, or the simulation of intelligence in computers or machines, was coined back in 1956, only a decade after the creation of the first electronic digital computers. (…) In the 2000's scientists were developing AIs that excelled in specific areas, such as winning at chess, cleaning the kitchen floor and recognising human speech. Such "narrow" AIs, as they are called, have superhuman capabilities, but only in their specific areas of dominance. A chess-playing AI cannot clean the floor or take you from point A to point B. Superintelligent AI (…) will combine a wide range of skills in one entity. More darkly, it might also use data generated by smartphone-toting (=*carrying*) humans to excel at social manipulation. (…)

Jann Tallinn, an Estonia-born computer programmer (…) became convinced that superintelligence could lead to an explosion or breakout of AI that could threaten human existence – that ultrasmart AIs will take our place on the evolutionary ladder and dominate us the way we now dominate apes. Or, worse yet, exterminate us.

(…) Tallinn began funding research into methods that might give humanity a way out: so-called friendly AI. That doesn't necessarily mean a machine or agent is particularly skilled at chatting about the weather, or that it remembers the names of your kids (…). It doesn't mean it is motivated by altruism or love. A common fallacy is assuming that AI has human urges and values. "Friendly" means something much more fundamental: that the machines of tomorrow will not wipe us out in their quest to attain their goals. (…)

Jaan Tallinn in 2013. Photograph: Michael Bowles/Rex/Shutterstock

Every AI, whether it's a Roomba or one of its potential world-dominating descendants, is driven by outcomes (= *results*). Programmers assign these goals, along with a series of rules on how to pursue them. Advanced AI wouldn't necessarily need to be given the goal of world domination in order to achieve it – it could just be accidental.

(…) People get overly preoccupied with what superintelligent AI is, Tallinn said. What form will it take? Should we worry about a single AI taking over, or an army of them? "From our perspective, the important thing is what AI does," he stressed. And that, he believes, may still be up to humans – for now.

Adapted from **https://www.theguardian.com/technology/2019/mar/28/can-we-stop-robots-outsmarting-humanity-artificial-intelligence-singularity**