The dangers of developing AI

Humans have imagined how the future will turn out for centuries. After it became clear that technology was in our world to stay, many of these imagined possible futures, have in some way incorporated technology into their predictions. This phenomenon has become its own genre in science fiction, and the works in this genre show how some might interpret the future during the time of the work’s writing. At the genre’s inception we see how technology can be used to reach new heights, for example in Jules Verne’s masterpieces “Around the World in Eighty Days” or “Journey to the Center of the Earth”. These show how Verne might have seen technology with the potential to help humans in scientific journeys, but the genre slowly evolved to show how technology would change the world for the worse. The works in dystopian science fiction usually show how technology creates massive conflicts, and many have shown machines becoming evil, thus making it their goal to eradicate all humans. And through these horrible depictions of our future, this has become one of the broad possibilities people think our future can take shape in. But according to Stuart Russel, pioneer in artificial intelligence, this isn’t a possibility to worry about, even though he does admit that the machines can cause us problems.

In the article “The rise of robots: forget evil AI – the real risk is far more insidious” from The Guardian, Russel comments on how he thinks artificial intelligence might hurt humans in the near future. He argues that the chance of an artificial intelligence turns downright evil, is fairly slim, but he does think that one might cause harm to humans unintentionally. “Someone building a self-driving car might instruct it never to go through a red light, but the machine might then hack into the traffic light control system so that all of the lights are changed to green”[[1]](#footnote-2) is one example given to how Russel thinks AIs might act. This action, while not breaking any of the rules the AI has been set, also minimizes the time for the car to reach its destination, which in most circumstances is what a self-driving car will be motivated to do. Because of problems like these, Russel argues that we need to train AIs to align with human values, as they will find loopholes in any work they will be doing, and without human values, they will act on these loopholes every single time.

In the video “Robots can predict the future … and so can you” from The Guardian, Anil Seth, professor at Sussex University, discuss how it might not be optimal to instill such human behavior in AIs. “In many context’s, human intelligence is actually a bit rubbish, and we make irrational decisions all the time”[[2]](#footnote-3), he explains, and remarks that AIs probably shouldn’t try to mimic human behavior, as that might put their enormous processing power to bad use, and in a single irrational decision, this power could be quite devastating. Alan Winfield, professor at Bristol Robotics Lab, also reminisces over the fact that AIs might experience hurt if they would be created with a conscience, and since we don’t understand one hundred percent how consciousness works, we might inadvertently put AIs through emotional hurt without the proper knowledge in that field.

Some of these conflicts go against one another, as on one hand, AIs would need to align with human values to work correctly how we want them to, but this could cause some of the problems discussed in the video, how the AI might act irrationally, if we teach it to function like a human, or that we might unintentionally cause it harm, if we aren’t cautious. This makes developing AIs hard, as the developer would have to weigh these moral options themselves, and if ever a fully functional general AI is developed, it might be used by its owner to do insidious deeds, just as any other powerful technology might be. AIs would simply be all the more devastating, as a general AI would be able to wreak havoc in every aspect of our lives, as they are general, and therefore able to apply its vast powers in any way it is required to.

Due to this, it seems like AI should be developed under numerous safety measures, and in a collaboration between multiple nations and/or institutes, as AIs could be used to make their owners politically and powerful, simply due to the military power an AI could possess, and thus it is needed to make sure that AIs are developed with all of mankind in mind, not simply as a tool for power and war.

In conclusion, we should not be scared of any sentient AI turning evil, though we do face some problems in developing one that would act accordingly to human values, and without this, an AI might bring unnecessary harm, simply while trying to optimize its work. Also, we should be aware that AIs will be able to cause humans harm, if it is developed to do so, and that because of this, no single state or organization should control any AIs alone, as humans tend to create conflicts, and with such a powerful device on any side of a battlefield, a lot of harm could be done.

Indledningstype brugt: Lidt en blanding mellem ”Etablér en historisk kontekst for dit emne”, og ”Fortæl en historie, der levendegør de pointer, din tekst skal diskutere”.

1. <https://www.theguardian.com/technology/2016/aug/30/rise-of-robots-evil-artificial-intelligence-uc-berkeley>, line 49-51. [↑](#footnote-ref-2)
2. <https://www.theguardian.com/lifeandstyle/video/2017/feb/02/robots-can-predict-the-future-and-so-can-you-video>, 2:38-2:45 [↑](#footnote-ref-3)