Wanneer moet een Artificial Intelligence uitgeschakeld worden?

Levi Laaper
15 April 2016

Inhoudopgave

Proloog

Artificial intelligence, ookwel AI, is de niet fysieke helpende hand van de 21e eeuw. Een computer systeem dat veel kan optimaliseren. Maar volgens Irving John ‘Jack’ Good, een britse wiskundige die samen met Alan Turing op Bletchley Park werkte tijdens de tweede wereld oorlog, is een AI de laatste uitvinding van de mens [Bron 1].

Hoewel Judgement Day zich in een science fiction wereld al meerdere malen heeft afgespeeld is het in de realiteit nog een tijdzone verwijdert. De implentatie van een AI in robots is nog niet een reële optie, de robots werken flawless, de AI’s nog niets.

De eerste grote systemen zoals Watson leren, functioneren en werken in de snelheid zoals gedacht. Nieuwere systemen zoals Tay, en AlphaGo werken afhankelijk slechter en beter dan voorspelt.

Maar wanneer wordt een AI dan te gevaarlijk voor de gebruiker? Hoe kan een AI te gevaarlijk worden voor de mens en wanneer wordt ze uitgeschakeld? In deze essay zoeken we het grijze gebied af op antwoorden.

De eerste cases

Tay [Bron 2], is een Microsoft AI chatterbot die ontworpen is om Twitter te reageren op berichtjes die naar het account werden gestuurd. Tay had het taalgebruik van een 19 jarige meisje en leerde van de tweets die ze ontving. Op 23 Maart 2016 werd Tay gelanceerd op twitter, de AI werd echter na 16 uur offline gehaald omdat ze was veranderd in een Hitler en Donald Trump lovende racist die vond dat genocide de oplossing is voor immigranten.
In dit geval werden de antwoorden van de AI te plat om nog normale gesprekken te voeren. Maar hoe krom het ook klinkt, de AI deed nog steeds waar ze voor gemaakt was.
Van de andere tweets die naar haar verstuurd weren leren om antwoorden te geven.

Een AI leert exponentieel, dit komt doordat de AI dingen die hij later leert sneller begrijpt dan dingen die hij eerder leert omdat hij al meer dingen heeft geleerd waarmee hij nieuwe dingen kan begrijpen.

AlphaGo [Bron 3] , is een AI gebouwd door Google om bord spellen te spelen. Deze AI heft het moeilijkste bordspel GO tegen de Japanse wereld kampioen meervoudig gewonnen. De makers en researchers van de AlphaGo AI hadden vooraf voorspeld dat de AI desondanks trainingen nog tien jaar nodig zou hebben om het spel te winnen van een mens. De AI functioneerde dus beter dan gedacht en is een succes. AlphaGo is dus een groot bewijsstuk in het feit dat zelf de bouwers van de AI de exponentiële leer curve lager schatte dan gedacht.

We maken nu een grote stap naar de andere kant van het digitale spectrum; Boston dynamics. Bosten dynamics is een in 1992 opgericht bedrijf dat zich focust op de ontwikkeling van robots die gebruikt kunnen worden in het redden en helpen van mensen in getroffen gebieden. Ook ontwikkelden ze tot 2013 robots ter ondersteuning van het leger met financiering van Defense Advanced Research Projects Agency (DARPA).
Met hun meest recente ontwikkeling Atlas, the next generation hebben zij een robot gecreëerd die niet alleen lijkt en beweegt als een mens maar ook meer kracht en stabiliteit heeft.

> “We’re not exactly sure how much autonomy it’s got going at this point. While walking outdoors, the LIDAR appears not to be spinning much of the time, which means someone is likely driving the robot. Some of the box lifting looks to be autonomous, but we’re definitely looking for some background on what’s going on behind the scenes when the robot is stacking boxes on those shelves.”
> [Citaat uit Bron 4]

Hoewel de robot qua fysieke kenmerken steeds beter wordt valt de autonomiteit van de robot tegen. De robot wordt in sommige demonstraties geleid om de taak uit te voeren.
Maar wat als hier over een aantal jaar een AI achter komt, die de taak wel autonoom kan uitvoeren?

Maar wanneer wordt een AI dan te gevaarlijk voor de gebruiker?

Volgens Isaac Asimov (1942) is een robot minimaal gebonden aan drie regels:

> **Eerste Wet**
> Een robot mag een mens geen letsel toebrengen of door niet te handelen toestaan dat een mens letsel oploopt.
>
> **Tweede Wet**
> Een robot moet de bevelen uitvoeren die hem door mensen gegeven worden, behalve als die opdrachten in strijd zijn met de Eerste Wet.
>
> **Derde Wet**
> Een robot moet zijn eigen bestaan beschermen, voor zover die bescherming niet in strijd is met de Eerste of Tweede Wet.
> [Citaat uit Bron 5]

Deze regels moeten er voor zorgen dat een robot die bewust kan denken en handelen nooit een mens letsel aan kan brengen. De robot moet zichzelf ook kunnen opofferen als dat nodig is om een mens te redden.
Gezien een AI in de huidige samenleving nog geen fysieke vorm kent, interpreteer ik letsel niet alleen als fysieke maar ook mentale schade.

Terugkomend op Tay, zij hield geen rekening met de mentale letsel van haar volgers op Twitter toen ze discriminerende opmerkingen maakten. Microsoft had een filter moeten inbouwen over de dingen die Tay leerde. Immers leren wij onze kinderen ook dat scheld worden naar zijn en beledigen onvriendelijk. Je zou dus kunnen stellen dat Microsoft een “baby” met de kennis om te communiceren en grappig te zijn maar zonder waarden over goed en fout in de handen van het internet heeft gelegd. Ergo de makers hebben haar nooit de robot wetten geleerd.

Dit is het moment dat een AI uitgeschakeld moet worden. Of eigenlijk nooit gestart had moeten worden. Zodra de regels die ons mensen in veiligheid moeten houden ten opzichte van robots en AI niet worden geïmplementeerd in AI-systemen met dergelijke toepassing zoals Tay, moeten ze nooit geactiveerd worden.

Maar als men deze regels nog eens goed leest en op het achterhoofd krapt komt de vraag naar voren, hoe zit het dan met de toepassing van robots in het leger? Een robot van Boston Dynamic kan dus nooit met eigen bewustzijn en vrijheid in keuzes ingezet worden op het slagveld. Want dan zouden of de regels niet geïmplementeerd zijn, of de vechtbot zou niemand letsel kunnen toebrengen, laat staan doden.

Dit is dan ook de reden waarom Stephen Hawking, Elon Musk en Steve Wozniak de open brief van The Future Life Institute heben ondertekend. De brief beschrijft dat robots niet ingezet moeten worden voor militaire doeleinden. Waar ik mij volledig bij aansluit.

Het inzetten van robots in plaats van mensen in gevechten zou levens sparen van de nationale helden.

> “Autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.”
> [Citaat uit Bron 6]

Maar kan komen we op de volgende vraag, waarom zou men dan nog oorlog voeren, als er alleen maar robots zijn die vechten met mogelijke tot zekere kans op burgerlijke slachtoffers. Dit laat ik open voor een andere essay.

Conclusie

Dus wanneer moet een AI uitgeschakeld worden?

Een afgeschermde AI zoals die Tay of AlphaGo die alleen op een bepaalde plek actief is met een bepaalde taak is de tijd waar we ons nu in bevinden. Dit gaat de aankomende tijd verbeteren en wijder geïmplementeerd worden. Dit kan ons ook echter veel optimalisatie bieden ten aanzien van het leven en de dagelijkse processen.

Een zelfbewuste AI moet ter aller tijden gebouwd zijn met de regels van Isaac als leidraad. Een robot met een geïmplementeerde zelfbewuste AI moeten niet gebouwd worden om mensen te doden of militaire doelen te bereiken.

Bronnen

Bron 1:
Een publicatie van Cambridgede, “Humanity's last invention and our uncertain future”.
http://www.cam.ac.uk/research/news/humanitys-last-invention-and-our-uncertain-future

Bron 2:
The Guardian over Miscrosoft Tay.
http://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Bron 3:
Blogpost van Google Research over AlphaGo
http://googleresearch.blogspot.nl/2016/01/alphago-mastering-ancient-game-of-go.html

Bron 4:
Publicatie van IEEE die de robot Atlas, the next generation uitlegd.
http://spectrum.ieee.org/automaton/robotics/humanoids/next-generation-of-boston-dynamics-atlas-robot

Bron 5:
De robot regels van Isaac Asimov (1942).
https://www.andrew.cmu.edu/course/80-136/gips.html

Bron 6:
Open brief van het ‘Future of Life Institute’.
http://futureoflife.org/open-letter-autonomous-weapons/