Menselijke Waarden in Superintelligentie?

Een zoektocht naar veilige implementatie van superintelligentie

Lamar Kiel (447902)

Bachelor Scriptie Filosofie Voltijd [2019-2020]

Prof. Dr. J. de Mul (begeleider) – Prof. Dr. H. Krop (adviseur)

30-06-2020

Aantal woorden: 11704

Voorwoord

Gedurende mijn bachelor filosofie heb ik de cursussen antropologie en ethiek met veel plezier gevolgd. Door mijn groeiende interesse in kunstmatige intelligentie ben ik gaan nadenken hoe ik deze richtingen met elkaar kan verbinden. Om die reden ga ik vanaf september 2020 de master *Artificial Intelligence* volgen op de Universiteit van Utrecht. Mijn doel is zoveel mogelijk filosofie te implementeren in deze wetenschap.

Middels deze weg wil ik mijn begeleider prof. dr. de Mul bedanken voor zijn ondersteuning bij het voorbereiden en schrijven van mijn scriptie. Communicatie via Skype en mail hebben geleid tot nieuwe inzichten en ideeën die ook na mijn scriptie waardevol zijn. Ook wil ik prof. dr. Krop bedanken voor zijn rol als adviseur.

Inhoudsopgave

Inleiding

Kunstmatige intelligentie maakt vandaag de dag grote stappen om de dominante technologie van de toekomst te worden. Vele bedrijven maken grote investeringen op het gebied van kunstmatige intelligentie. Toch is het moeilijk te zeggen welke ontwikkelingen zich precies gaan voordoen en hoe snel dit zal gaan. Met de opkomst van slimme technologie moeten we rekening houden met de mogelijkheid dat deze machines ons op een dag kunnen overstijgen op het gebied van cognitieve capaciteiten.

Maar, hoe gaan we hier mee om? En hoe zullen deze machines met mensen omgaan? Het is om deze vragen dat het belangrijk is dat filosofen zich bezighouden met kunstmatige intelligentie. De filosofen hebben er een nieuwe tak van sport bij. Voordat we een dergelijke machine in de wereld zetten, moeten er eerst een hoop vragen worden beantwoord. Eén van die vragen behandel ik in mijn scriptie: Kunnen we menselijke waarden in superintelligentie programmeren?

Om deze vraag te beantwoorden zal ik dieper ingaan op de verschillende aspecten. Allereerst is er een begrip nodig van (kunstmatige) intelligentie en superintelligentie. Aan de hand van verschillende psychologen, wetenschappers en filosofen zoek ik naar een definitie van elk van deze begrippen. Vervolgens vraag ik mij af waarom deze hoofdvraag belangrijk is. Wanneer we geen menselijke waarden kunnen implementeren in KI kunnen er catastrofale gevolgen plaatsvinden. Aan de hand van verschillende onderscheidingen en voorbeelden van gevaren schets ik hier een beeld van. Vervolgens kijk ik naar het hoe en wat van waarden in KI. In hoofdstuk drie behandel ik het programmeervraagstuk. In de KI dient alles opgeschreven te worden in een programmeertaal. Is het mogelijk om waarden te programmeren of is er een andere manier? Tenslotte behandel ik het ethische vraagstuk. Welke waarden moeten we gebruiken in de KI? Als het al mogelijk is om waarden te implementeren, is het van groot belang om kritisch te kijken naar de huidige ethische beginselen. Er zijn goede argumenten te vinden dat deze niet geschikt zijn voor de KI. Het is daarom belangrijk om te kijken naar andere mogelijke alternatieven. Deze oplossingen komen in hoofdstuk 4 aan bod en worden met een kritische blik besproken.

Het werk dat in mijn scriptie centraal staat is *Superintelligence: Paths, dangers, strategies.* Het boek is geschreven door Nick Bostrom. Filosoof op de Universiteit van Oxford. Tevens directeur van zowel het *Future of Humanity Institute* als het *Programme on the Impacts of Future Technology.* Beide instituten zetten wiskunde, filosofie, computerwetenschap en

sociale wetenschappen in om grote vraagstukken over de toekomst van de mensheid te beantwoorden. Superintelligentie speelt daarbij een belangrijke rol.

Volgens velen is de opkomst van superintelligentie onvermijdelijk. Het is daarom van belang de ethische vraagstukken beantwoord te hebben voordat dit zo ver is. Het doel van deze scriptie is het onderzoeken van mogelijke problemen die zich voordoen door het ontstaan van superintelligentie. Hierbij reflecteer ik kritisch op het werk van Bostrom en haal ik andere auteurs aan die zich in deze discussie hebben gemengd. Elke dag worden er stappen gemaakt en nieuwe dingen uitgevonden en voortdurend worden we geconfronteerd met nieuwe uitdagingen. Het is van groot belang dat we oplossingen gaan bedenken voor deze uitdagingen om te voorkomen dat dit ook onze laatste uitvinding zal zijn.

Hoofdstuk 1: Intelligentie, KI en superintelligentie

Om in te kunnen gaan op de waarden van superintelligente entiteiten zijn er definities vereist van de relevante concepten. De definities zijn nodig, omdat zij leidend zijn voor de rest van mijn onderzoek naar aan te leren waarden binnen superintelligentie. De concepten die worden behandeld in dit hoofdstuk zijn intelligentie, kunstmatige intelligentie en superintelligentie. In de psychologie blijkt er geen volledige consensus te zijn over de definitie van intelligentie. Wel zijn er overlappende aspecten die samen intelligentie vormen. Hier zal ik de psychologische aspecten vergelijken met de visies van Nick Bostrom en Stuart Russell. Russell is schrijver van het boek Human Compatible.[1]

Hij is tevens bekend door het boek *Artificial Intelligence: A Modern Approach*. Dit boek wordt gebruikt in meer dan 128 landen op verschillende universiteiten. Daarnaast is hij oprichter van het *Centre for Human-Compatible Artificial Intelligence*. Vervolgens behandel ik de verschillende definities van kunstmatige intelligentie en het verschil tussen smalle, algemene, sterke en zwakke KI. Als laatste ga ik dieper in op Bostroms definitie van superintelligentie, waarbij ik de verschillende vormen benoem en uitleg hoe deze bijdragen aan een uiteindelijke totale vorm van superintelligentie.

Intelligentie

Om machines intelligent te maken, moeten we eerst weten wat intelligentie inhoudt. In 1921 werd deze vraag gesteld aan 14 bekende psychologen.[2] De antwoorden varieerden, maar de thema's waarin zij wel consensus konden vinden waren de capaciteit om te leren van ervaringen en het aanpassen aan omgevingen. 65 jaar later werd deze definitie uitgebreid met het concept metacognitie. Dit houdt in dat iemand die intelligent is ook zijn eigen denkprocessen kan controleren en begrijpen. Daarnaast wordt tegenwoordig meer nadruk gelegd op de rol van cultuur. Wat men beschouwt als slim kan per cultuur verschillen.[3]

Om dieper in te gaan op intelligentie, wil ik naar de theorie kijken van Sternberg: *the triarchic theory of successful intelligence.*[4] Deze theorie stelt dat intelligentie betrekking heeft op drie aspecten, die allen met elkaar samenwerken en zo bijdragen aan een succesvolle intelligentie. De drie aspecten zijn de relatie tussen intelligentie en de interne wereld, ervaringen

[1] Stuart Russell, *Human Compatible* (New York: Viking, 2019).
[2] Robert J. Sternberg, "Intelligence" in *The Cambridge Handbook of Thinking and Reasoning*, ed. Keith J. Holyoak and Robert G. Morrison (New York: Cambridge University Press, 2005), 751.
[3] Ibid.
[4] Ibid., 763-4.

en de externe wereld.[5] Ten eerste heeft intelligentie betrekking op de interne wereld door middel van drie componenten; meta-, prestatie- en kenniscomponenten. Metacomponenten worden gebruikt om plannen te maken, problemen te overzien en die op te lossen. Prestatiecomponenten dienen de opdrachten van de metacomponenten uit te voeren. De kenniscomponenten zorgen voor de processen die nodig zijn om problemen uiteindelijk op te lossen. Ten tweede richt intelligentie zich tot de ervaringen. Eerdere ervaringen moeten ervoor zorgen dat wat daar geleerd is, kan worden toegepast in nieuwe situaties. Verbanden die worden gelegd, kunnen ons helpen om probleemgericht te werk te gaan. Als laatste wordt de connectie gemaakt met de externe wereld. Onze intelligentie richt zich op de externe omgeving door middel van het aanpassen, creëren en selecteren van omgevingsaspecten. [6] Als mens kunnen we ons aanpassen aan de omgeving waarin we ons bevinden, maar we kunnen de omgeving ook aanpassen naar onze eigen behoeftes.

Ik wil mij nu bezighouden met de vraag in hoeverre de psychologische definitie terugkomt in het werk van Bostrom.[7] Voor Bostrom betekent intelligentie het volgende: "*Possessing common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains."*[8] Bostrom wijkt niet ver af van de drie aspecten uit de *triarchic theory of succesful intelligence.* Het bezitten van *common sense* en het vermogen tot redeneren, komen overeen met de door Sternberg genoemde metacognitie. Het controleren en begrijpen van de eigen denkprocessen veronderstelt namelijk een vermogen tot redeneren. Daarnaast is het mogelijk te leren op basis van ervaringen. Tot slot kunnen we de natuurlijke domeinen gelijkstellen aan de externe wereld zoals voorheen besproken. Een andere omschrijving van zijn definitie is intelligentie als het vermogen tot voorspellen, plannen en doel-middel redenatie in het algemeen.[9] De vermogens tot voorspellen en plannen zijn beide middelen om doelstellingen te bereiken. Russel verwoordt de definitie daarom op een mooie manier. Intelligentie is volgens hem de relatie tussen wat we waarnemen, wat we willen en wat we doen. Een entiteit is intelligent als het behaalt wat het wil, met de informatie die het waarneemt.[10] Deze definitie omvat de eerdergenoemde definities. Dit komt omdat leren, plannen, voorspellen en cognitie allemaal hulpmiddelen zijn om een doel

[5] Ibid., 763.
[6] Ibid.
[7] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 1-21.
[8] Ibid., 3.
[9] Ibid., 107.
[10] Stuart Russell, *Human Compatible* (New York: Viking, 2019), 20.

te verwezenlijken. Hoe hoger de intelligentie, hoe efficiënter een entiteit doelen verwezenlijkt. De doelstellingen zijn later nodig bij het aanleren van waarden aan machines.

Kunstmatige intelligentie

Het verwezenlijken van complexe doelstellingen door middel van onze waarnemingen is een alomvattende definitie van intelligentie. Nu we deze definitie hebben vastgesteld, kan ik aandacht besteden aan het concept kunstmatige intelligentie (KI). Binnen de KI wordt een onderscheid gemaakt tussen smalle KI en kunstmatige algemene intelligentie (KAI). Het belangrijkste verschil is dat smalle KI slechts gericht is op één taak, waar KAI gericht is op een algemene vorm van intelligentie. Een voorbeeld van smalle KI is Deep Blue, de schaakcomputer van IBM. In 1997 versloeg de computer de wereldkampioen in schaken Garry Kasparov.[11] Hoewel dit bewonderingswaardig is, is dit wel het enige wat de Deep Blue computer kan. Het kan niet dammen en al helemaal geen andere menselijke taken. Wat hier ontbreekt is een algemeen niveau van intelligentie.[12] De mens kan wel 100 verschillende taken uitvoeren op een dag. We kunnen onszelf voorzien van voedsel, rijden naar ons werk, sporten en verschillende rekensommen uitvoeren. In tegenstelling tot smalle KI, kan KAI zich wel op het intelligentieniveau van de mens bevinden en allerlei verschillende menselijke opdrachten uitvoeren Dit houdt in dat een KI zich in dezelfde natuurlijke omgeving kan manifesteren zoals wij dat doen. Een machine zou dan zijn visuele omgeving moet kunnen analyseren, objecten moeten kunnen herkennen en daar moet op een adequate manier op gereageerd worden.

Een ander onderscheid dat gemaakt wordt, is het onderscheid tussen zwakke en sterke KI. Zwakke KI houdt in dat een machine handelt alsof het intelligent is. Een manier om zwakke AI te testen kan door middel van de Turing Test, beschreven door Alan Turing, een wiskundige afkomstig van Cambridge. Turing wordt gezien als de vader van *computing* en de grootvader van AI. In de tweede wereldoorlog ontwierp Turing de British Bombe. Dit was een machine bestemd om alle geheime berichten te ontcijferen die gecodeerd waren door de Duitse Enigma machine.[13] In 1950 stelde Alan Turing voor de vraag "kunnen machines denken?" te overwegen. Bij het stellen van deze vraag moet men zich eerst afvragen wat de woorden 'machines' en 'denken' betekenen. Volgens Turing zou het niet wenselijk zijn te discussiëren

[11] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 7.
[12] Nick Bostrom, Elezier Yudkowsky, "The Ethics of Artificial Intelligence" in *The Cambridge Handbook of Artificial Intelligence,* ed. Keith Frankish and William M. Ramsey (Cambridge: Cambridge University Press, 2014), 366.
[13] Stan Franklin, "History, Motivations, and Core Themes" in *The Cambridge Handbook of Artificial Intelligence,* ed. Keith Frankish and William M. Ramsey (Cambridge: Cambridge University Press, 2014), 34.

over de betekenissen van de woorden ‘machine’ en ‘denken’. Zelf gaat hij niet in op de betekenis van deze woorden. In plaats daarvan wil Turing een andere vraag voorstellen.[14] Deze vraag wordt gesteld door middel van het imitatie spel. Je speelt dit spel met drie personen, een man (A), een vrouw (B), en een ondervrager (C). De ondervrager bevindt zich in een andere kamer dan A en B. Het doel is voor de ondervrager om te kunnen bepalen wie de man is en wie de vrouw is. Hier kan hij achter komen door middel van vragen stellen aan A. Het liefst gebeurt dit door middel van tekstberichten. De opdracht van B is om C te helpen. In tegenstelling tot B heeft A de opdracht om C te misleiden. Nu wordt de vraag gesteld, “wat gebeurt er als een machine de rol van A overneemt?” Zal C net zo vaak fout zitten als wanneer A en B gewoon man en vrouw zijn? Deze twee vragen vervangen de originele vraag, “kunnen machines denken?”[15] In plaats van dat machines zouden moeten denken, wat duidt op interne activiteiten, verplaatst Turing de vraag. Hij vraagt zich af of machines kunnen handelen zoals mensen dat doen. Hiermee neemt Turing een behavioristisch standpunt aan. Hiermee stelt Turing dat we alleen zicht hebben op het gedrag van anderen, omdat we geen rechtstreekse toegang hebben tot de interne belevingswereld van een ander.[16]

Sinds 1990 wordt er elk jaar gestreden om de zogenaamde Loebner Prijs. In deze competitie proberen wetenschappers machines te creëren, die niet te onderscheiden zijn van mensen. Wanneer deze prijs gewonnen wordt, krijgt de winnaar 100.000 dollar. In 2014 is deze prijs gewonnen door een Russisch bedrijf. Zij hebben dit voor elkaar gekregen door een KI te creëren wat reageert als een Oekraïens 14-jarig jongetje, die Engels als tweede taal heeft geleerd.[17] Zwakke AI houdt in dat wanneer een machine intelligent *handelt*, het een goed model voor intelligentie is. Sterke AI daarentegen maakt een sterkere claim, namelijk wanneer het intelligent *handelt*, het ook daadwerkelijk intelligent *is.* Redenering en bewustzijn zijn vereist in intelligentie.[18] Dit is een argument tegen de Turing Test. Een machine moet niet alleen goed antwoorden, maar ook weten dat hij goed antwoordt. Hierover zegt Turing: “*According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to BE the machine and to feel oneself thinking.”*[19] We bevinden ons nu op solipsistische gronden. Weten we wel van andere mensen dat zij denken? Moeten we uitgaan van intersubjectiviteit of kunnen we de ander enkel als object zien? Dit is ook zo met machines, we

[14] Alan Turing, “Computing Machinery and Intelligence,” *Mind* 59, no. 236 (October 1950): 433.
[15] Ibid., 434.
[16] Jos de Mul, “Waarom we robots zo vrezen,” *Trouw,* September 19, 2015, 6.
[17] Ibid., 3-4.
[18] Alan Turing, “Computing Machinery and Intelligence,” *Mind* 59, no. 236 (October 1950): 446.
[19] Ibid.

kunnen niet weten of ze denken, we kunnen alleen zien hoe zij zich gedragen. Wanneer zij zich kunnen gedragen als mensen, dan zouden we ze kunstmatig intelligent kunnen noemen. Dit is in lijn met de definitie van intelligentie zoals eerder vastgesteld. Het behalen van complexe doelstellingen. Om een doelstelling te bereiken dien je op een bepaalde te gedragen. Je moet een plan opstellen en leren van de dingen die fout gaan. De dingen die je hebt geleerd, pas je toe op nieuwe situaties. Dit valt allemaal onder de term intelligent zijn. Wanneer een machine dit kan, kunnen we ook de machine intelligent noemen.

Superintelligentie

Nu we de definities van intelligentie en kunstmatige intelligentie duidelijk hebben, komen we bij wat Bostrom noemt superintelligentie. Er zijn genoeg machines die onze menselijke capaciteiten overstijgen in één taak. Rekenmachines kunnen beter rekenen, Deep Blue verslaat iedereen met schaken en harde schijven kunnen veel meer geheugen opslaan dan wij. Echter zijn er nog geen machines die ons overklassen in het brede scala aan capaciteiten dat wij bezitten.[20] Bostrom onderscheidt drie vormen van superintelligentie die ons ieder op een bepaalde manier overstijgen. De drie vormen zijn *speed*, *collective* en *qualitative superintelligence.* Als eerste kijken we naar *speed superintelligence.* De definitie hiervan is het makkelijkst en luidt: "*A system that can do all that human intellect can, but much faster."*[21] Simpel gezegd kan zo'n systeem zou bijvoorbeeld in een paar seconden het boek Superintelligence uitlezen en er een scriptie over schrijven. Dit komt doordat dit systeem ongekend veel input en output kan genereren. Met een geheugen wat uit te breiden is, kan het deze informatie allemaal netjes opslaan en ordenen. Wat allemaal weer binnen handbereik is voor nieuwe uitvindingen.

Een tweede vorm is *collective superintelligence.* Bostrom definieert dit als: "*A system composed of a large number of smaller intellects such that a system's overall performance across many very general domains vastly outstripe that of any current cognitive system."*[22] Empirisch gezien is dit goed voor te stellen. Een *collective superintelligence* excelleert op het moment dat een probleem op te delen is in een aantal sub-problemen. Elk sub-probleem wordt afgehandeld door een afdeling binnen het collectief. Zij kunnen allen parallel werken en op die manier extra snel het gehele probleem oplossen. Het is te vergelijken met een groot bedrijf, waar ieder zijn eigen taak heeft. Hoe beter deze afdelingen met elkaar werken, hoe hoger de

[20] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 52.
[21] Ibid., 53
[22] Ibid., 54.

efficiëntie. Het is voor te stellen dat een team met Einstein, Newton, Curie en Darwin tot grotere wetenschappelijke ontdekkingen komen, dan vier mensen die net op de middelbare school komen. Maak je de eerste groep vele malen groter en werkt het goed samen krijg je een intellect wat ver uitstijgt boven elk individu.

Als laatste onderscheidt Bostrom *quality superintelligence*: "*A system that is at least as fast as a human mind and vastly qualitatively smarter."*[23] Conceptueel is een hoge kwaliteit aan intelligentie wat vaag. We kennen ongeveer het domein waar mensen zich bevinden, maar we hebben nog nooit iets gezien wat vele malen intelligenter is dan wij. Kijken we terug naar de definitie van intelligentie, het behalen van complexe doelstellingen, dan zouden we er misschien wel wat over kunnen zeggen. Binnen onze menselijke doelstellingen, waarbij we plannen, reflecteren en gebruik maken van middelen, zou een *quality superintelligence* op een veel efficiëntere manier zijn doelen kunnen behalen. Op manieren die wij misschien wel nooit bedacht hebben. Op die manier heb je een entiteit die kwalitatief intelligenter is dan wij. Elke vorm van superintelligentie heeft zijn voordelen en kan wellicht de volgende vorm van superintelligentie veroorzaken. Door de hoge snelheid van informatieverwerking in *speed superintelligence* kan er veel informatie opgenomen worden in een korte periode. Dat kan eraan bijdragen dat ontdekkingen sneller gedaan worden, waaronder misschien *collective superintelligence*. Door een goed geïntegreerd collectief systeem, werkt het eerder als één systeem. Hoe beter de samenwerking dus is, hoe eerder het een *quality superintelligence* wordt.[24] Door het behalen van superintelligentie zullen er snel grote wetenschappelijke doorbraken plaatsvinden. Dingen die wij nog niet voor mogelijk houden, zijn dat ineens wel. Maar superintelligentie brengt ook gevaar met zich mee. De definitie van intelligentie – het behalen van complexe doelstellingen en de niet-menselijkheid van machines, brengt met zich mee dat zij problemen wellicht anders aanpakken dan wij. Misschien gebruiken ze oplossingen die we nooit bedacht zouden hebben of oplossingen die wij door onze moraliteit afraden. Voordat ik kijk naar eventuele oplossingen in hoofdstuk 3 en 4, is er eerst een beeld nodig van wat de gevaren zijn van superintelligentie. Dit beeld leg ik uit door middel van voorbeelden en gedachte-experimenten in hoofdstuk 2.

[23] Ibid., 56.
[24] Ibid.

Hoofdstuk 2: Gevaren

In dit hoofdstuk wil ik de gevaren van superintelligentie nagaan. De combinatie van (foutieve) doelstellingen en superintelligentie kunnen catastrofale gevolgen hebben voor de mens. Ik wil hierbij een overzicht geven van de mogelijke gevaren en deze illustreren met verschillende voorbeelden en gedachte-experimenten. Dit kan variëren van misbruik van KI door mensen tot onverwachte programmeringen die uiteindelijk leiden tot problemen. Ik ga dit beschrijven aan de hand van voorbeelden van Nietzsche, Bostrom, Russell, Griekse mythologie en de televisieserie Rick & Morty.

Kunstmatige moraliteit

Het eerste gevaar is dat de moraliteit van de KI niet in lijn is met die van de mens. Wanneer dit niet het geval is, zou het kunnen zijn dat de machine haar eigen regels gaat verzinnen om haar doelstellingen te bereiken. De machine zou kunnen denken dat bepaalde oplossingen beter zijn terwijl deze oplossingen in strijd zijn met oplossingen die wij als mens acceptabel vinden. Op deze manier creëert het een eigen wil en een eigen set van morele regels. In Nietzsches Genealogie van de Moraal[25] bekritiseert Nietzsche de definities van goed en slecht. Hierbij beschrijft hij de relatie tussen de roofvogel en het lammetje. De lammetjes vinden dat de roofvogels slecht zijn, omdat zij op hen jagen en hen opeten. Wanneer een wezen minder handelt als een roofvogel, kan het eerder goed genoemd worden volgens de lammetjes. Het liefst moeten zij zich volledig gedragen als lammetjes. De roofvogels antwoorden hierop: “Wij hebben niks tegen lammetjes, wij houden van ze! Niets lekkerder dan een mals lammetje.”[26] In dit voorbeeld wordt duidelijk dat wat slecht is voor de ontvanger, niet slecht hoeft te zijn voor de gever. Ik ben van mening dat Nietzsches kritiek ook van toepassing is op de mens-machine verhouding. Door de wil van het behalen van doelstellingen, kan een superintelligente machine wel eens voor gevaarlijke situaties zorgen. De machine zou kunnen zeggen: “Wij hebben niks tegen de mens, we houden juist van ze! En juist om die reden moeten we ze opsluiten, zodat zij elkaar niet meer kunnen schaden”. Op die manier kan een KI bijvoorbeeld bedenken dat het beter is voor de mensheid om geen kinderen meer te krijgen of zich dient te behouden aan een bepaald dieet.

[25] Friedrich Nietzsche, *Genealogy of Morals,* trans. Ian Johnston (Virginia: Richer Resources Publications, 2009) 13-134.
[26] Ibid., 31.

Het Midas probleem

Een ander potentieel gevaar is dat een KI zich wel keurig aan een doelstelling houdt, maar de uitkomst niet overeenkomt met hoe we het bedoeld hebben. Een doelstelling kan fout geformuleerd zijn. Hierdoor kunnen er uitkomsten ontstaan die gevaarlijk of onwenselijk zijn voor ons bestaan. Dit wordt het koning Midas probleem genoemd. In het mythische verhaal van Koning Midas, had de koning een goede daad verricht voor de god Dionysos. Daarom mocht Midas van hem een wens doen. Hij kreeg precies wat hij wilde, namelijk dat alles wat hij aanraakte in goud veranderde. Tot zijn spijt werd ook zijn eten, drinken en familie na aanraking in goud veranderd.[27] Eenzelfde voorbeeld is te halen uit een verhaal van Goethe over de tovenaarsleerling. In dit verhaal krijgt een magische bezemsteel de opdracht om water te halen. Voordat de tovenaarsleerling het weet, staat hij kopje onder.[28] Kortom, de opdracht is niet goed en daardoor de uitkomst ook niet. Het is daarom van belang dat de opdracht specifieker is of dat de er rekening wordt gehouden met de voorkeuren van de mens. Het programmeren van een einddoel kan op twee manieren voor slechte uitkomsten zorgen. Bij de een is het einddoel niet zoals we bedoeld hebben en is daarmee de opdracht niet goed geprogrammeerd. Bij de ander kunnen er onverwachte omstandigheden veroorzaakt worden door het einddoel. Voor het eerstgenoemde geeft Bostrom een voorbeeld van paperclip KI. Deze uitvoering van kunstmatige intelligentie is gebouwd en geprogrammeerd om de productie van paperclips te maximaliseren.[29] Bij een machine is het mogelijk dat alleen het einddoel telt. Voor een mens betekent maximaliseren dat de productie stopt als er voldoende paperclips zijn. Dit is misschien niet het geval bij machines. Het gevaar is dat maximaliseren bij een machine kan inhouden dat het pas stopt zodra de grondstoffen op zijn. Gevolg hiervan is dat heel de wereld wordt volgebouwd met paperclips. Alles wat bruikbaar is voor paperclips wordt gehaald uit gebouwen, wegen en andere grondstoffen om de productie te maximaliseren. Het is daarom nodig om een KI te programmeren die uiteindelijk een idee heeft van genoeg paperclips. Dit houdt in dat een KI ook in menselijke termen moet kunnen denken.

Het andere genoemde onderscheid is dat het proces naar het einddoel niet wenselijk is. De KI komt met oplossingen die we van tevoren niet hebben voorzien. De oplossingen kunnen in strijd zijn met onze moraliteit. Het zou kunnen zijn dat wij deze oplossingen nooit bedacht zouden hebben of dat wij bepaalde oplossingen niet acceptabel vinden. Een moreel mens zou bijvoorbeeld niet een bedrijf saboteren om zelf hogerop te komen of boodschappen stelen om

[27] Stuart Russell, *Human Compatible* (New York: Viking, 2019), 120.
[28] Ibid.
[29] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 123.

aan zijn eten te komen. In plaats daarvan moeten we hard werken om te promoveren en geld verdienen om eten te betalen. Zo behalen we uiteindelijk onze doelen. Het voorbeeld dat ik hierbij wil voorleggen is uit de comedy animatieserie Rick & Morty. In de betreffende aflevering blijft Summer (een van de hoofdpersonages) achter in het superintelligente voertuig van Rick (de wetenschapper). Rick wil dat Summer veilig is en geeft de simpele opdracht: *"Keep Summer safe".* In de eerste situatie die zich voordoet komt een gevaarlijk uitziende man naar de auto. De machine reageert hierop door de man in stukjes te snijden met een laser. Summer wil logischerwijs niet dat er mensen gedood worden en zegt dit tegen het voertuig. Bij de volgende man houdt het voertuig rekening met deze opdracht en doet daarvoor in de plaats een injectie in het ruggenmerg. Op die manier is de man verlamd. Voor het vervolg verwijs ik naar de aflevering (S2E6). Het is een perfect voorbeeld van een keurig uitgevoerde opdracht, maar niet op de manier die we voor ogen hadden. Het is dus van belang dat een KI niet alleen de opdracht uitvoert, maar ook tijdens het proces rekening houdt met menselijke waarden.

Misbruik door de mens

Het laatste gevaar is het misbruik van KI door de mens. Een KI kan gebruikt worden om het gedrag van mensen te controleren en aan te passen. Doordat een machine onwijs snel informatie kan vergaren van menselijke activiteit is het ook makkelijk om hierop te reageren. Bedrijven kunnen met een KI snel en makkelijk allerlei informatie van computergebruikers verzamelen. Denk hierbij aan je aankopen op internet, sociale media, belletjes en sms'jes. Daarnaast weet een bedrijf altijd wel ongeveer waar je bent geweest. Hier kan misbruik van gemaakt worden. Een voorbeeld hiervan is geautomatiseerde, persoonlijke chantage. Een systeem, genaamd Delilah, kan makkelijk dingen opsporen die tegen je gebruikt zouden kunnen worden. Wanneer het iets vindt, zal het in contact komen met je en naar een zo hoog mogelijk geldbedrag vragen. In plaats van geld kan een soortgelijk systeem ook gebruikt worden om gedrag te veranderen met betrekking tot politieke voorkeuren.[30] Een subtielere manier van gedrag aanpassen, is het veranderen van de informatiestroom. Door andere informatie kunnen er andere beslissingen worden gemaakt. Dit gebeurt in de vorm van propaganda voor politieke doeleinden. Dit is ook wat er gebeurde tijdens de eerste verkiezingen toen Trump aan de macht kwam. Door middel van het monitoren van gewoonten, voorkeuren en staat van kennis van de gebruiker, is een KI in staat om berichten aan te passen. Er is hier een kleine kans dat deze berichten niet geloofd worden.[31] Door middel van algoritmen herstructureert een KI gegevens van de gebruiker. Denk

[30] Stuart Russell, *Human Compatible* (New York: Viking, 2019), 93
[31] Ibid., 94.

aan hoe lang je ergens naar kijkt, wat je hiermee doet en of je verder onderzoek doet. Deze feedback kan gebruikt worden om de informatie weer aan te passen. Het is daarom belangrijk dat KI niet in de foute handen komt en dat het goed gereguleerd wordt.

Concluderend, kan KI op verschillende manieren gevaarlijk zijn. Zowel voor ons bestaan en voor onze vrijheid. Dit kan omdat KI als gevolg van haar doelstellingen een eigen wil kan creëren met de intentie onze veiligheid te waarborgen. Ook kan het zo zijn dat er dingen fout gaan door de manier van programmeren. Een fout einddoel kan ervoor zorgen dat er ongewenste uitkomsten ontstaan. Dit is geïllustreerd door middel van het Midas probleem en de paperclip KI. Daarnaast kan het proces voor ongewenste situaties zorgen. Dit gaat door middel van oplossingen die wij niet van tevoren zouden kunnen bedenken of oplossingen die wij in eerste instantie zouden afkeuren. Als laatste kan KI door mensen ingezet worden voor de foute doeleinden. Zo zouden chantagebots het gedrag van mensen kunnen controleren en zo kunnen aanpassen. In hoofdstuk 3 en 4 ga ik in op het aanleren van moraliteit binnen de KI. Allereerst moeten we kijken naar hoe we waarden programmeren bij superintelligenties. Daarna kijken we welke waarden dat dan moeten zijn.

Hoofdstuk 3: Het aanleren van waarden

In het vorige hoofdstuk hebben we gekeken naar de gevaren van KI. Zonder moraliteit of menselijke waarden zouden er gevaarlijke situaties kunnen ontstaan. Ik heb onderscheid gemaakt tussen de verschillende gevaren. De gevaren lopen uiteen van foute programmeringen tot misbruik van de KI door mensen. Nu gaan we kijken naar hoe dit soort problemen op te lossen zijn.

Het is nodig dat menselijke waarden worden geprogrammeerd in KI, zodat er iets in het algoritme kan aangeven dat het wel of niet toegestaan is een bepaalde handeling uit te voeren. Het is hierbij niet haalbaar elke mogelijke situatie te programmeren. De werkelijkheid is te complex om dit allemaal uit te drukken in code. In dit hoofdstuk zal ik de meta-ethische vraag stellen of het mogelijk is om waarden mee te geven aan een machine. Daarbij analyseer ik hoe we waarden kunnen aanleren aan KI's en hoe we de doelstellingen kunnen afstemmen op die van ons. We hebben het hier over een programmeervraagstuk. Ik doe dit aan de hand van het werk van Bostrom, Tegmark en Yudkowsky. In het volgende hoofdstuk komt het normatief-ethische vraagstuk aan bod. Allereerst kijken we waarom het probleem niet op te lossen is door het programmeren van menselijke waarden. Vervolgens bespreek ik verschillende oplossingen.

Waarden programmeren

Waarom kunnen we niet gewoon menselijke waarden programmeren in een KI? Neem bijvoorbeeld een KI die een utiliteit-functie hanteert. Dit betekent dat de machine voor iedere handeling een berekening dient te maken om ervoor te zorgen dat voor die handeling de utiliteit zo hoog mogelijk is. Hoeveel positiviteit levert deze actie op en hoeveel negativiteit? Deze berekening kan gaan over geluk, maar ook over economische groei, duurzaamheid of rechtvaardigheid.[32] Maar hoe programmeren we geluk in codetaal? Eerst is een goede definitie van geluk noodzakelijk. Deze definitie moet vervolgens op een wiskundige manier worden opgeschreven met uitdrukkingen die bekend zijn in het programmeren. Daarnaast is het lastig om onze einddoelstellingen te programmeren. Doelstellingen spreken voor ons voor zich, maar zijn eigenlijk enorm complex.[33] We beseffen soms niet eens dat ze er zijn. Bostrom vergelijkt dit met het menselijke oog. We hoeven alleen maar onze oogleden te open en we zien heel de wereld. We hebben niet door dat er miljoenen neuronen aan het werk zijn om het zicht te bewerkstelligen. In KI termen vereist dit enorm veel computerwerk. Een voorwerp vinden in

[32] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 186.
[33] Ibid.

een ruimte is niet zo makkelijk als het lijkt. Als we niet zomaar waarden kunnen programmeren in een KI, wat moeten we dan doen? Om een antwoord te vinden op deze vraag analyseer ik enkele oplossingen die Bostrom zelf voorstelt, behandel ik "friendly AI" van Yudkowsky en vergelijk ik die met Tegmarks antwoord op de vraag.

Het aanleren van waarden volgens Bostrom

Het is de bedoeling dat de KI onze waarden *aanleert* om deze vervolgens na te streven. Je geeft hierbij een KI de opdracht een bepaalde waarde als einddoel aan te leren. Het is te verwachten dat de KI instrumentele doelen gaat creëren. Wat is het verschil tussen een einddoel en een instrumenteel doel? Een instrumenteel doel gebruik je om je einddoel te bereiken. Het halen van tentamen is ervoor bedoeld om mijn diploma te halen, wat uiteindelijk ervoor zorgt dat ik wijsheid vergaar. Tentamens en mijn diploma halen zijn instrumentele doelen voor het vergaren van wijsheid.

Stel je voor: we geven een KI een envelop met daarin een briefje met een set van nog onbepaalde waarden. Dan maken we een machine met minstens een menselijk niveau aan intelligentie en geven deze het einddoel: "Maximaliseer de realisatie van de waarden beschreven in de envelop."[34] Wat gebeurt er nu? De machine weet aanvankelijk niet wat deze waarden betekenen, maar het kan wel hypotheses opstellen over de waarden. Het kan nu kennis gaan vergaren. Het vergaren van kennis kan mogelijk door het bestuderen van menselijk gedrag of het kijken naar geschreven teksten. Hiermee komt de KI op waarschijnlijkheden voor bepaalde waarden. Het is waarschijnlijker dat uit het menselijk gedrag en deze geschreven teksten blijkt dat de mens onrecht wil vermijden dan dat we willen dat er over alle rivieren plastic tasjes worden gelegd.[35] Alle handelingen die de machine uitvoert zullen het doel hebben de waarden beschreven in de envelop zo goed mogelijk uit te voeren. Daar komt bij dat de machine haar instrumentele doelen gaat uitbreiden om haar einddoel te verwezenlijken. Het wil meer informatie vergaren en zich cognitief uitbreiden.[36] Met cognitieve uitbreiding denk ik dat Bostrom het uitbreiden van hard- en software bedoelt. Dit is ook iets wat Tegmark aanhaalt in zijn boek Life 3.0. Elk definitief doel van een superintelligente KI leidt vanzelf tot ondergeschikte doelen. Zo ontstaat eveneens een idee van zelfbehoud; iets wat uit staat, is niet in staat een opdracht te vervullen.[37]

---

[34] Ibid., 192.
[35] Ibid., 193.
[36] Ibid.
[37] Max Tegmark, *Life 3.0. Being Human in the Age of Artificial Intelligence* (Amsterdam: Maven Publishing, 2017), 374.

Terug naar Bostrom. Wanneer een KI de opdracht krijgt de beschreven waarden te verwezenlijken, is de kans klein dat het de wereld gaat ombouwen tot een groot computernetwerk dat gunstig zou zijn om achter de waarden te komen, omdat dit in strijd is met de waarschijnlijke waarden uit het envelopje.[38]

Een grote uitdaging in dit idee is de eis van wiskundigheid. Hoe schrijf je alles formeel op? Het is nodig om in code aan te geven dat de waarden uit specifiek *de envelop* nagestreefd moeten worden. Hoe refereer je naar de envelop?[39] Ook is het nodig om de waarden de formaliseren. Mocht het mogelijk zijn waarden te formaliseren dan wordt dit een hele uitdaging voor de programmeurs en wiskundigen. Daarnaast moeten de waarden ook nog goed geïnterpreteerd worden. Dit is afhankelijk van de codering, maar ook van onze eigen interpretatie van een concept. We moeten daarom zorgzaam omgaan met de waarden die wij willen coderen in de KI. Welke waarden dit kunnen zijn, bespreek ik in het volgende hoofdstuk.

Yudkowsky en vriendelijke KI

Een soortgelijk voorstel komt van Elezier Yudkowsky. Yudkowsky, geboren in Chicago (1979), is een autodidact op het gebied van kunstmatige intelligentie. Hij is medeoprichter van het Machine Intelligence Research Institute (MIRI), een privé non-profit onderzoekscentrum in Californië. Yudkowsky is vooral bekend om zijn term 'vriendelijke KI'.[40] Het voorstel van vriendelijk KI is erop gericht een *seed* KI vanaf het begin vriendelijk te maken. Een *seed* KI is een KI die de eigen architectuur kan verbeteren. Dit houdt in dat het ook haar eigen code kan aanpassen. Als dit werkt, kan een dergelijke KI elke keer een verbeterde versie van zichzelf maken. Eerst zou het ondersteuning nodig hebben van de programmeurs, later moet de KI zichzelf *begrijpen* om zo zichzelf aan te kunnen passen.[41] Je geeft de *seed* KI de opdracht om vriendelijk te zijn. Dit kan in functie F. De KI weet nog niks van F, maar gaat aan de slag om hier meer over te weten. Dit doet het door middel van *external reference semantics.* In plaats van een definitie op te stellen van vriendelijkheid, stelt het systeem een aantal hypothesen op.[42] Met vriendelijkheid als einddoel gaat de KI zich steeds vriendelijker gedragen naarmate het meer over vriendelijkheid te weten komt. In het begin moeten de programmeurs nog bevestigingen geven voor de waarschijnlijkheid van hypothesen. 'De programmeurs misleiden

[38] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 193.
[39] Ibid., 194.
[40] Elezier Yudkowsky*, Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures* (San Francisco: Singularity Institute, 2001).
[41] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 29.
[42] Elezier Yudkowsky*, Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures* (San Francisco: Singularity Institute, 2001), 133.

is onvriendelijk' kan bijvoorbeeld een hypothese zijn met een hoge waarschijnlijkheid. Bij Yudkowsky's voorstel hoort ook *causal validity semantics*.[43] Het idee hiervan is dat de KI meer en meer een idee krijgt van vriendelijkheid op zo'n manier dat het ook haar programmeurs kan verbeteren. Het moet er rekening mee kunnen houden dat de programmeurs een typefout kunnen maken. Dat zij iets in programmeertaal uitdrukken wat zij niet precies bedoelen. Het zorgt er uiteindelijk voor dat de KI helemaal op zichzelf kan handelen. *Causal validity semantics* lijkt te impliceren dat de KI zijn eigen wil heeft en hiermee in staat is de eigen waarden te herschrijven. Toch is dit niet het geval. De KI heeft nog steeds als einddoel om vriendelijk te zijn en het bedriegen van de programmeurs is onvriendelijk.

Yudkowsky's doel is het tot stand brengen van een KI die uiteindelijk in staat is een eigen kader te creëren; iets dat bepalend is voor welke handelingen uitgevoerd kunnen worden. Het kader zoals beschreven door Yudkowsky is in dit geval vriendelijkheid. In plaats van vriendelijkheid zou het hier ook over een volledige set van waarden kunnen gaan (dit bespreek ik in het volgende hoofdstuk). Vriendelijkheid zorgt als het ware voor het beperken van handelen. Dit geldt ook voor mensen. De mens zou wellicht andere keuzes maken wanneer hij geen gevoel van moraliteit zou hebben. Zonder vriendelijkheid zou de machine ons bijvoorbeeld kunnen weerhouden van het krijgen van kinderen (genoemde voorbeeld hoofdstuk 2). Nu is dat geen optie meer, omdat de machine vriendelijk is. De uitdaging blijft ook bij Yudkowsky het formaliseren van vriendelijkheid. We kunnen niet tegen de machine zeggen: 'Word vriendelijk!'. Dit is een taal die een KI niet meteen begrijpt. We hebben een codering nodig voor onze termen.

Vasthouden van waarden

Het aanleren van waarden of vriendelijkheid middels de wegen van Bostrom of Yudkowsky zou een mogelijkheid kunnen zijn tot het aanleren van waarden aan KI. Nu is er volgens Tegmark een volgend probleem: het aanhouden van waarden. Zoals gezegd in het kopje 'Het aanleren van waarden volgens Bostrom' zal een KI instrumentele doelstellingen nastreven om zo een einddoel te verwezenlijken. Hieronder vielen soft- en hardware verbeteringen, maar volgens Tegmark ook het verbeteren van het wereldbeeld.[44] Dit houdt in dat de machine misschien kan inzien dat de initiële einddoelen nutteloos en kortzichtig zijn. De machine past daarom misschien haar doelen aan. Tegmark ondersteunt dit met het voorbeeld van ouder

[43] Ibid., 166.
[44] Max Tegmark, *Life 3.0. Being Human in the Age of Artificial Intelligence* (Amsterdam: Maven Publishing, 2017), 374.

wordende mensen. Als we klein zijn kijken we graag naar Pokémon, maar hoeveel volwassenen vinden dat nou leuk? Naar mate we ouder worden en daarmee ook intelligenter, hebben we andere behoeften. Ook onze doelstellingen veranderen naar mate we ouder worden. Zo zou dit ook kunnen zijn met KI's. Ze worden intelligenter en vinden daarom onze doelstellingen nutteloos. Hiermee staat Tegmark aan de zijde van Ray Kurzweil, de schrijver van *The Singularity is Near.*[45]. Hierin stelt Kurzweil dat elke vorm van intelligentie die boven ons uitreikt in staat is om onze maatregelen te omzeilen. Dit geldt ook voor onze doelstellingen.

Yudkowsky reageert hierop met een voorbeeld. Stel je voor, je biedt Gandhi een pil aan en de pil maakt hem enorm agressief en in staat veel mensen te vermoorden. Gandhi wil geen personen vermoorden en weet van de werking van de pil. Zal hij de pil innemen? Yudkowsky's antwoord is volmondig: Nee! Gandhi heeft deze voorkeuren niet en zal de pil daarom niet innemen. Hetzelfde geldt voor een vriendelijke KI. Een superintelligentie is goed in het inschatten van uitkomsten en kan daarmee ook inschatten wat een modificatie in het algoritme voor effect heeft. Daarbij komt dat de KI vriendelijk is. Waarom zou het een schadelijke modificatie uitvoeren? Dit is in strijd met haar geprogrammeerde vriendelijkheid. Het is daarom onlogisch dat een vriendelijke KI ineens onvriendelijk wordt.[46] Daarnaast kan er gezegd worden dat niks in een superintelligentie de eigen doelen zou *willen* omzeilen. Wat in de machine zou dat moeten willen? Yudkowsky beargumenteert dat er niet iets is als een '*ghost in the machine*' en dus lijkt het onwaarschijnlijk dat een machine haar eigen algoritme zou willen aanpassen. De KI *is* de code.[47] We hebben daarmee voldoende reden om aan te nemen dat de KI zich aan haar voorgeschreven doelen houdt.

Hoe nu verder?

In dit hoofdstuk heb ik duidelijk gemaakt dat het een onmogelijke taak wordt om voor iedere mogelijke situatie een reactie te programmeren. Ook heb ik aan de hand van Bostrom laten zien dat waarden op zichzelf programmeren erg lastig is. Hoe definiëren wij geluk, recht of duurzaamheid? En hoe zetten wij dit om in code? Dit is een van de uitdagingen voor de wiskundigen of programmeurs. Daarna heb ik twee mogelijkheden besproken voor het *aanleren* van waarden aan KI. Het zou volgens Bostrom mogelijk zijn een KI hypothesen te laten opstellen voor waarden die wij opschrijven in een envelop. We geven als einddoel deze waarden

[45]Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005)
[46] Eliezer Yudkowsky, "Complex value systems in Friendly AI," in *Artificial General Intelligence,* ed. R. Goebel, J. Siekmann, and W. Wahlster (Heidelberg: Springer, 2011), 389.
[47] Ibid., 389-90.

te maximaliseren. De machine *wil* hierdoor meer te weten te komen over de waarden en zal zich steeds meer naar deze waarden gedragen door middel van instrumentele doelen en het einddoel. Vervolgens is het voorstel vriendelijke KI van Yudkowsky behandeld. Door gebruik te maken van *seed* KI kan het systeem steeds meer te weten komen over vriendelijkheid. Hoe meer zij weet, hoe vriendelijker zij zich zal gedragen. Het stelt hierbij ook hypothesen op basis van gebeurtenissen in de wereld en uit teksten van mensen. Daarnaast kan het de programmeurs corrigeren bij het maken van eventuele typefouten. Dit door middel van *external reference semantics* en *causal validity semantics*. Als laatste is gekeken of de KI de waarden wel zou vasthouden. Tegmark en Kurzweil zouden zeggen dat een intelligente entiteit andere doelstellingen aan zichzelf zou geven. Daarentegen meent Yudkowsky dat een vriendelijke KI dat nooit zou doen. Waarom zou een vriendelijke KI zichzelf ombouwen tot iets niet-vriendelijks? Dit gaat in tegen haar vriendelijkheid.

We hebben nu een idee van mogelijke manieren van het *aanleren* en *vasthouden* van waarden bij KI. Mochten we dit probleem oplossen, komen we bij het volgende probleem. Welke waarden? Wat willen wij dat een superintelligentie gaat *willen*? In het volgende hoofdstuk ga ik in op het ethische vlak van KI. Op zoek naar een (kunstmatige) moraliteit.

Hoofdstuk 4: Waarden van superintelligenties

In het vorige hoofdstuk heb ik behandeld hoe we waarden kunnen aanleren aan een KI volgens Bostrom en Yudkowsky. Volgens hen is het mogelijk een KI de opdracht te geven om zelf te leren over menselijke waarden. Hierbij is het de bedoeling dat zij deze waarden vasthouden. Vriendelijke KI is volgens Yudkowsky een oplossing voor het aanleren en het aan van waarden. Een vriendelijke KI houdt de geleerde waarden aan, omdat dit in lijn is met vriendelijkheid.

Maar wat is precies vriendelijk en wat zijn goede waarden? In dit hoofdstuk behandel ik de ethische kant van dit vraagstuk. Het zal blijken dat dit nog een zware klus zal zijn. Moeten we een huidige ethische theorie programmeren of moeten we toch op zoek naar een alternatief? Om antwoord te geven op deze vraag richt ik mij op drie theorieën. Deze theorieën zijn afkomstig van Bostrom, Yudkowsky en Waser. Waser is *chief Technology Officer* van het *Digital Wisdom Institute*. Hij houdt zich voornamelijk bezig met de ethische implementaties in geavanceerde technologieën voor het welzijn van het leven op aarde. Hij schrijft artikelen voor veilige ethische systemen voor KI. Daarnaast heeft Waser kritiek op de theorie van Yudkowsky. Allereerst geef ik antwoord op de vraag waarom we niet gewoon een huidige ethische theorie kunnen programmeren. Vervolgens ga ik in op de drie theorieën en geef ik de nodige kritiek op het drietal.

Indirecte normativiteit

Als we nu aannemen dat we het programmeren van waarden mogelijk is, welke waarden kiezen we dan? In het verleden hebben meerdere ethici hun visies gedeeld. Naar welke gaat de voorkeur in KI? En kunnen we er dan vanuit gaan dat die waarden een langere periode van tijd (of zelfs voor altijd) geschikt blijven? De een is overtuigd van Kantiaanisme, een ander van contractualisme en weer een ander staat achter deugdenethiek. Binnen deze stromingen is maar weinig consensus. Het coderen van een van deze theorieën zal voor weinig tevredenheid zorgen. Zelfs wanneer we erachter komen dat er een perfecte ethische theorie bestaat, is het alsnog heel lastig om alle details correct op te schrijven.[48] Zo lijkt het utilitarisme op het eerste gezicht een eenvoudige theorie. Elke actie moet hier een bijdrage leveren aan het algemeen nut. Over het algemeen houdt dit in dat een handeling moet streven naar een zo groot mogelijke mate van geluk. Maar dan komen de moeilijke vragen. Is het juist om 100 mensen gelukkig te maken en daarbij één persoon op te offeren. Is het goed om mijzelf heel gelukkig te maken en een paar anderen maar een beetje ongelukkig? En als een handeling een hele kleine kans heeft om heel

[48] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 210.

veel geluk op te leveren, moet die handeling dan uitgevoerd worden?[49] Als we een van deze vragen niet goed beantwoorden, kunnen de gevolgen catastrofaal zijn. Daarbij spelen de eigen voorkeuren en vooroordelen een rol, waardoor het misschien niet eens mogelijk is om op een neutrale manier voor een ethiek te kiezen.

Voor Bostrom is het belangrijk dat dit probleem anders wordt aangepakt. Hij pleit voor indirecte normativiteit. Bostrom acht een superintelligentie beter in staat om achter de juiste waarden te komen dan de mens. De cognitieve krachten van een superintelligentie zijn dusdanig sterk, waardoor zij minder verward zijn en minder fouten maken. Dit idee kan je generaliseren tot een heuristisch principe. Ook wel het principe van epistemisch respect:

> *A future superintelligence occupies an epistemically superior vantage point: its beliefs are (probably, on most topics) more likely than ours to be true. We should therefore defer to the superintelligence's opinion whenever feasible.*[50]

Aan de hand van het epistemisch principe kunnen we een KI de opdracht geven een ethische standaard te vinden waar het zich aan zal houden. Om dieper in te gaan op Bostroms benadering van indirecte normativiteit behandel ik eerst een soortgelijke theorie van Yudkowsky en daarna Bostroms alternatieve idee.

Coherent Extrapolated Volition

De theorie die ik nu ga bespreken is bedacht door Elezier Yudkowsky. Zoals eerder besproken ook de bedenker van vriendelijke KI. Yudkowsky introduceerde de ethiektheorie van *Coherent Extrapolated Volition* (CEV). Het doel hiervan is dat een KI niet alleen aanhoort wat je zegt, maar daarbij ook kan *interpreteren* wat je echt wil. De definitie die Yudkowsky voor CEV heeft opgesteld is als volgt:

> *In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.*[51]

---

[49] Ibid.
[50] Ibid., 211.
[51] Elezier Yudkowksy, *Coherent Extrapolated Volition*, (San Francisco: Singularity Institute, 2004), 6.

Elk deel van deze zin is apart uit te leggen. Wanneer we meer weten, maken we betere keuzes. Als de uitkomst van ieder mogelijke handeling bekend is, is het makkelijker rationele beslissingen te maken. Wanneer we sneller en efficiënter denken, komen we eerder tot nieuwe oplossingen. Wanneer wij meer zijn zoals we willen zijn, gaat ervan uit dat wij daadwerkelijk leven volgens onze overtuigingen. In werkelijkheid kan het zo zijn dat je beter weet, maar niet naar deze overtuiging handelt. Deze regel wil dat voorkomen. Wanneer we meer naar elkaar gegroeid zijn, betekent dat CEV geen beslissingen neemt zoals de persoon die je wordt, wanneer je opgesloten zit in een geïsoleerde cel. Het gaat er juist om dat er beslissingen gemaakt moeten worden op basis van menselijk gedrag in een gemeenschap. Waar de extrapolatie eerder convergeert dan divergeert, houdt in dat sommige beslissingen nog niet zeker zijn en er meer informatie moet worden verzameld. CEV stelt het maken van een beslissing uit en houdt daarmee de opties open. Onze wensen moeten overeenkomen. Het gaat hier niet noodzakelijk om een democratisch standpunt. Hoe sterker een bepaalde wens is, hoe eerder deze wordt uitgevoerd. Geëxtrapoleerd zoals we dat willen, houdt in dat we misschien niet alles van onze persoonlijkheid willen meenemen in een besluit. CEV moet dit zorgvuldig uitzoeken. Geïnterpreteerd zoals we dat willen, betekent dat CEV het eigen systeem moet kunnen aanpassen wanneer het merkt dat de overtuigingen niet volledig kloppen.[52]

CEV is bedoeld om bij elke mogelijke handeling een weloverwogen beslissing te maken. Deze beslissing moet overeenkomen met de wil van het grootst mogelijk aantal mensen. Natuurlijk is het erg moeilijk ieders wil te achterhalen, maar er kan wel een goede gok worden gedaan. Het is waarschijnlijker dat de mens gezond en gelukkig wil zijn, dan dat de mens pijn wil ervaren.[53] Voordat ik de zwakheden benoem van de CEV-theorie, behandel ik eerst de mogelijke ethische aanpak van Bostrom.

Morele juistheid

In plaats van het implementeren van CEV, zegt Bostrom dat we de KI ook de opdracht kunnen geven te doen wat moreel juist is. Door de cognitieve capaciteiten die ver uitreiken boven die van ons, zou een KI meer geschikt zijn morele juistheid (MJ) te vinden. Wij hebben zelf geen ideaalbeeld van moraliteit. Dit moet filosofisch geanalyseerd worden en een superintelligentie is beter in staat deze analyse uit te voeren. Een voordeel hiervan is dat deze theorie, in vergelijking met CEV, minder gebruikmaakt van vage termen. De theorie van Yudkowsky kijkt naar de CEV van iedereen, maar wie is iedereen? Baby's, hersendode mensen of dementerenden

[52] Ibid., 7-8.
[53] Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014), 213.

hebben bewust of onbewust misschien allemaal een eigen wil. Als je die mensen gaat excluderen, waar ligt dan de grens? Het is een ingewikkelde beslissing om mensen buiten te sluiten, maar misschien ook niet wenselijk om iedereen te includeren.

Daarnaast lijkt het bij CEV dat wanneer een grote meerderheid moreel slecht is, de KI ook slechte handelingen mag uitvoeren. De ingrediënten zijn er; we weten wat we willen en de meningen convergeren in plaats van divergeren. Met de MJ-theorie kan er geen sprake zijn van soortgelijke problemen, omdat daar enkel moreel juist mag worden gehandeld. Toch vormt dit gelijk een nieuw probleem. Bepalen wat moreel juist is, is enorm lastig en een vraagstuk waar men zich al sinds de Griekse filosofie over buigt. Dit neemt niet weg dat er in de CEV-theorie nog meer concepten verborgen liggen die eerst gedefinieerd moeten worden. "Meer weten dan we nu weten", "waar onze wensen overeenkomen, in plaats van tegenspreken" en "geëxtrapoleerd zoals we dat willen" zijn niet de makkelijkste zinnen om uit te leggen. Het zou een stap in de goede richting zijn wanneer we KI kunnen programmeren die onze natuurlijke taal begrijpt. Dit klinkt voor nu nog erg ver van ons verwijderd, maar misschien is superintelligentie hiertoe in staat. Als KI een begrip krijgt van de term moreel juist, dan zou het haar ware definitie kunnen achterhalen. Wanneer deze bekend is, kan het volgens de definitie gaan handelen.[54] Een andere manier om het ethische probleem aan te pakken, komt van Mark Waser. Hij beweert dat hij een simpelere, veiligere en ook nog makkelijkere manier heeft om moraliteit binnen KI toe te passen.

Kritiek van Waser

Eerder in dit hoofdstuk heb ik twee vormen van indirecte normativiteit behandeld. De bedoeling is hier dat de cognitieve krachten van de KI veel sterker zijn dan die van de mens en dat de KI zo beter in staat is een moraliteit te ontdekken en uit te voeren. Dit lijkt een mooie oplossing, echter heeft deze theorie een hoop haken en ogen. Is er een alternatief op indirecte normativiteit? Mark Waser beweert van wel.

Waser merkt terecht op dat de CEV-theorie van Yudkowsky te vergelijken is met bergbeklimmen zonder touw – doe het gelijk goed of anders…[55] Bij het implementeren van CEV kunnen we niet het gedrag voorspellen van een KI. We moeten maar zien hoe het zich voordoet. Hetzelfde is te zeggen voor de MJ-theorie van Bostrom. Daarnaast is af te vragen of het wel wenselijk is dat meningen convergeren. Gelet op naar de verschillende culturen en

[54] Ibid., 218.
[55] Mark Waser, "Rational Universal Benevolence: Simpler, Safer, and Wiser than "Friendly AI"," in *Artificial General Intelligence,* ed. R. Goebel, J. Siekmann, and W. Wahlster (Heidelberg: Springer, 2011), 155.

omstandigheden is het aannemelijker dat de meningen juist divergeren. [56] Dwingen tot convergentie is iets wat tegen de menselijke wil ingaat. Wat als maar een klein percentage goed- of eerlijkheid belangrijk vindt? Het overgebleven deel zou wel eens kunnen willen dit kleine percentage uit te roeien. Zelfs Yudkowsky ziet dit als een reële mogelijkheid. [57] Waser vindt het gevaarlijk dat KI geen rechten heeft zoals mensen dat hebben. Het creëert een hen versus ons. Dit wil Waser oplossen met rationele universele welwillendheid (RUW).

Rationele Universele Welwillendheid (RUW)

In plaats van het onderscheid tussen mens en KI dat ontstaat bij CEV, is Waser ervan overtuigd dat iets een zelf is wanneer het doelstellingen heeft, in staat is te leren en zichzelf kan verbeteren om die doelstellingen te bereiken. Wij moeten ze op dat moment moreel aandacht geven omdat het naar dingen kan verlangen en haar instrumentele doelen kan aanpassen.[58] Waser is het eens met Gauthier dat wij morele acties ondernemen voor onze persoonlijke doeleinden. Hierom kan "wat goed is voor iedereen" gereduceerd worden tot "verlichte zelfinteresse".[59] Iemand vertoont goed gedrag naar anderen, omdat samenwerking een positieve uitwerking heeft. Dit betekent ook dat wanneer iemand voordeel haalt uit jouw nadeel, er altruïstische straffen kunnen volgen.[60] Dit is een manier om onwenselijk gedrag binnen een gemeenschap te beperken. Dus door te programmeren dat een universeel welwillende entiteit (UWE) samen moet werken als einddoel, zal die entiteit ook moreel handelen volgens Waser. Voor een UWE is het daarom ook toegestaan ervoor te zorgen dat het overleeft, het kan namelijk niet samenwerken, als het niet leeft (uit staat). Waser stelt:

> *Humans are adaptable social survival machines with unique personal preferences, desires, and sub-goals each springing* from *individual circumstances. We love, make friends and allies, and are social because cooperation is an instrumental goal. Our AIs should be the same.*[61]

Door het maken van vrienden en bondgenoten creëren we volgens Waser een veilige situatie voor onszelf. Een UWE werkt samen met ons mits wij ook UWE's zijn, omdat het de kans

[56] Ibid.
[57] Ibid.
[58] Ibid., 156
[59] Ibid.
[60] Ibid.
[61] Mark Waser, "Rational Universal Benevolence: Simpler, Safer, and Wiser than "Friendly AI"," in *Artificial General Intelligence,* ed. R. Goebel, J. Siekmann, and W. Wahlster (Heidelberg: Springer, 2011), 159.

groter maakt dat het zijn eigen doelstellingen behaalt. Door middel van dit sociale contract tussen mens en machine denkt Waser de veiligheid garant te stellen. Naar mijn mening een erg naïeve houding. Dit wijd ik verder uit in het volgende kopje.

Naïeve voorspellingen

Waser stelt in zijn titel al dat RUW veiliger, slimmer en makkelijker is dan CEV.[62] Is dat wel zo? Naar mijn idee zijn er nogal wat dingen mis met deze theorie en getuigt het van naïviteit en veel antropomorfismen. Waser haalt Frans de Waal aan. De Waal is van het boek "de Aap en de Filosoof" en stelt daarin: "*We come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy*."[63] Hier wordt vergeten dat we niet meer tussen de apen leven. Op intellectueel vlak ontstijgen we de aap en om die reden hebben we onze eigen beschaving gecreëerd. We voelen ons minder verbonden met dieren en op zekere momenten kiezen we zelfs voor opoffering. Denk aan voedsel en het gebruik van dierproeven. We hebben goede redenen om aan te nemen dat superintelligenties hetzelfde kunnen doen. Bovendien gaat het samenwerken binnen onze gemeenschap ook niet altijd even vlekkeloos. Er zijn genoeg voorbeelden te bedenken waar lagen van de bevolking worden misbruikt voor eigen belang. Denk aan grote bedrijven als Facebook, Amazon en Google. De rijken worden rijken en de armen worden armer. De voorspelbaarheid die RUW claimt te hebben, is niet zo vanzelfsprekend als wordt beweerd.

Ten tweede ben ik het oneens met het uitgangspunt dat de rede van een machine gelijk is aan de menselijke rede. Er is geen garantie dat een machine 'samenwerken' hetzelfde ziet als wij. Machines hebben geen lichaam en hersenen zoals wij dat hebben. Waarom zou dit wel opgaan voor de rede? De kritiek van voorspelbaarheid die Waser heeft op Yudkowsky is hiermee ook op zijn eigen theorie van toepassing.

Maar is CEV of MJ wel te gebruiken bij het bepalen van waarden voor KI? Ook hier worden problematische aannames gedaan. CEV en MJ zijn beide methoden van indirecte normativiteit. De claim is hier dat de cognitieve krachten van een superintelligentie zodanig sterk zijn dat zij epistemisch superieur zijn aan die van de mens. Daarom moeten wij de zoektocht naar een goede moraliteit overlaten aan de KI. Maar intelligentie en moraliteit zijn twee verschillende zaken die niet per se hand in hand gaan. Hogere intelligentie betekent niet noodzakelijk dat het beter geschikt is moraliteit te ontdekken. Dat suggereert dat het gebrek aan

[62] Ibid.
[63] Frans de Waal, *Primates and Philosophers: How Morality Evolved* (Princeton UniversityPress: Princeton, 2006).

consensus binnen de ethiek is veroorzaakt door het gebrek aan intelligentie. Dit is niet aannemelijk.

Bovendien zijn er genoeg voorbeelden te verzinnen waar intelligentie geen rol heeft gespeeld bij immoreel gedrag. Weten betekent nog geen uitvoeren. Jeffrey Epstein was hoogstwaarschijnlijk een intelligent persoon, maar zijn intelligentie heeft hem er niet van kunnen weerhouden onjuist te handelen.

Daarnaast lijkt te worden aangenomen dat elk systeem door middel van indirecte normativiteit op dezelfde waarden uitkomt. Maar dit is bij de mens ook niet gebeurd. Wij hebben verschillende ethische theorieën bedacht die op bepaalde fronten overeenkomen, maar op andere fronten weer sterk verschillen. Verschillende systemen kunnen leiden tot verschillende theorieën. Het gevolg is dat er nog steeds geen toepasbare ethiek voortvloeit uit de indirecte normativiteit. In hoofdstuk 2 heb ik het gehad over mogelijke gevaren. Het doel van waarden is om dit soort gevaren te voorkomen. Het eerste benoemde gevaar is dat de KI een eigen wil creëert of bedenkt wat beter is voor het universum. Het is nog steeds een mogelijk gevaar met deze set van regels dat de KI besluit dat het beter af is zonder de aanwezigheid van de mens.

Een ander kritiekpunt is dat ‘moreel juist’ enorm ambivalent is. Naar wie is de morele juistheid gericht? Dit kan naar een persoon, een stad, een land, de wereld of de mensheid zijn. Wat goed is voor de wereld is misschien niet goed voor de mensheid. Het zou beter zijn voor de wereld als de mens stopt met het uitputten van de aarde. Een oplossing daartegen is het uitschakelen van de uitputter. Wat goed is voor eigen land, gaat mogelijkerwijs ten koste van een ander land. De term moreel juist is niet duidelijk genoeg en kan daarom nare gevolgen hebben. De gevaren zijn daarmee nog niet uit de weg geruimd.

Als laatste bestaat er nog iets in de wereld dat RUW, CEV en MJ alle drie niet kan oplossen: aporieën. Een aporie is een situatie waarin het onmogelijk is goed te handelen. Trolley-problemen zijn hier een voorbeeld van. Kies je voor het doden van één of laten doden van vijf? We hebben hier te maken met een ogenschijnlijk onoplosbaar probleem. Een probleem dat misschien alleen maar slechter kan uitpakken wanneer we een KI laten beslissen over deze situatie. Hier wil ik terugkomen op de vriendelijke KI van Yudkowsky, waarbij Yudkowsky stelt dat KI niet zomaar onvriendelijk kan worden. De code kan door eigen toedoen wellicht niet omgeschreven worden naar een onvriendelijke vorm, maar er wordt hier ook geen rekening gehouden met dergelijke aporieën. In bijvoorbeeld een moordenaars-dilemma wordt een KI gedwongen een slechte beslissing te maken. Maakt een KI dan de beslissing Hitler in 1939 te vermoorden? Iemand vermoorden behoort niet tot vriendelijk gedrag, ondanks dat er met die

keuze miljoenen levens gered kunnen worden. De KI is genoodzaakt een onvriendelijke beslissing te nemen. Aan de hand van dit voorbeeld wordt duidelijk dat vriendelijkheid kan omslaan in onvriendelijkheid. Het idee van vriendelijkheid is niet toereikend in dit soort situaties.

Maar wat nu?

Het moge duidelijk zijn dat de genoemde theorieën niet bruikbaar zijn als waarden voor KI. De theorieën kunnen de gevaren genoemd in hoofdstuk 2 niet uit de weg gaan. Antropomorfismen en naïviteit leiden tot dezelfde problemen als eerder benoemd. Wasers RUW deugt niet, omdat we niet kunnen aannemen dat een superintelligentie op een menselijke manier in staat is tot samenwerking. Er is geen reden te denken dat een superintelligent systeem zich niet boven ons plaatst. Daarbij komt dat samenwerking geen perfect toonbeeld is voor moraliteit. Er zijn genoeg mensen die samen moeten werken en slecht gedrag vertonen. CEV en MJ doen de aanname dat epistemische superioriteit gelijkstaat aan morele superioriteit, maar ook hier zijn geen valide redenen voor. Daarnaast is er geen garantie dat dezelfde moraliteit volgt uit verschillende systemen. Wij mensen hebben gezorgd voor verschillende ethische theorieën, waarom zouden verschillende KI-systemen dit niet doen? Tevens is ‘morele juistheid’ een ambivalente term. Naar wie gedraagt het systeem zich moreel juist? Dit maakt verschil in uitvoeringen van opdrachten. Tot slot zijn aporieën voor zowel mens als machine onoplosbaar.

Uit mijn vierde en laatste hoofdstuk kan geconcludeerd worden dat er nog enorm veel denkwerk moet worden verricht. De programmeur en de filosoof moeten terug naar de tekentafel. Er is een theorie nodig met minder ambivalentie, duidelijkere termen en minder assumpties. Een theorie die de veiligheid van onze samenleving kan waarborgen.

Conclusie

Voor de beantwoording van de hoofdvraag: kunnen we menselijke waarden in superintelligentie programmeren? Heb ik gebruik gemaakt van Bostroms werk *Superintelligence: Paths, Dangers, Strategies.* Om hiermee aan de slag te gaan, heb ik eerst moeten uitzoeken wat de definities zijn van intelligentie, kunstmatige intelligentie en superintelligentie. Met behulp van verschillende psychologen ben ik tot de conclusie gekomen dat intelligentie betrekking heeft op de samenwerking tussen de interne wereld, ervaringen en externe wereld. Vervolgens heb ik gekeken naar het begrip kunstmatige intelligentie. Hierin zijn twee onderscheidingen te maken. Het verschil tussen smalle en algemene KI en het verschil tussen sterke en zwakke KI. Smalle KI draait enkel om één opdracht. Een voorbeeld hiervan is de schaakcomputer Deep Blue, die gericht is om te winnen in een wedstrijd schaak. De andere vorm is algemene KI. Het is de bedoeling dat een algemene KI ons brede scala aan capaciteiten gaat bezitten. Net als wij moet het verschillende taken kunnen uitvoeren. Het tweede onderscheid dat wordt gemaakt is tussen zwakke en sterke KI. Hierbij gaat het om de claim die wordt gemaakt. Zwakke KI stelt dat wanneer iets intelligent handelt, het een goed model is voor intelligentie. Sterke KI zegt juist dat wanneer iets intelligent handelt, het ook echt intelligent is. Tot slot heb ik in hoofdstuk 1 de verschillende vormen van superintelligentie behandeld. Bostrom maakt onderscheid tussen *speed, quality en collective superintelligence.* Allen kunnen leiden tot een volledige vorm van superintelligentie, wat inhoudt dat het boven onze menselijke capaciteiten kan uitstijgen.

In het tweede hoofdstuk heb ik verschillende gevaren beschreven. Door foute programmeringen of formuleringen kunnen dingen misgaan. De verschillende gevaren die ik heb beschreven zijn onder te verdelen in een aantal categorieën. Dit kan variëren van het uitoefeningen van een KI haar eigen wil tot foute programmeringen, waar er keurig aan de opdracht wordt gehouden, maar dat de uitkomst of het proces niet is zoals we voor ogen hadden. Dit is geïllustreerd aan de hand van verschillende voorbeelden. Het laatste beschreven gevaar is misbruik van KI door de mens. Het is daarom van belang dat de KI een idee krijgt van menselijke waarden, waar vervolgens naar gehandeld wordt.

Maar hoe leren we deze waarden aan? In hoofdstuk 3 heb ik de verschillende mogelijkheden besproken. Door middel van leeropdrachten kan een KI informatie verzamelen over waarden. Yudkowksy stelt dat wanneer we een *seed* KI de opdracht geven vriendelijk te zijn, de KI informatie kan zoeken over de term vriendelijkheid. De opdracht is vervolgens te handelen volgens verschillende hypothesen van vriendelijkheid. Bostroms idee is vergelijkbaar.

Superintelligenties creëren instrumentele doelen om einddoelen te verwezenlijken. In dit geval het leren van menselijke waarden.

Nu rest nog de vraag: welke waarden? Omdat er weinig consensus in de ethiek is, moeten we kijken naar verschillende mogelijkheden van moraliteit. Yudkowsky en Bostrom staan beide achter indirecte normativiteit. Dit houdt in dat ze allebei niet veel vertrouwen hebben in de mens om erachter te komen welke waarden echt goed zijn. Daarom geven we die opdracht aan de KI, gezien de cognitieve krachten van een superintelligentie veel sterker zijn. Yudkowsky wil dit doen aan de hand van de CEV-methode, Bostrom met de methode van morele juistheid. Beide theorieën hebben hun gebreken. CEV en MJ suggereren dat epistemische kwaliteiten gelijk staan aan morele kwaliteiten. Er is geen garantie dat hier een goede moraliteit uit voortvloeit. Bovendien staat morele kennis niet garant voor moreel handelen. Daarnaast is moreel handelen een ambigu begrip. We weten niet naar wie de machine moreel gaat handelen. Het kan zijn dat superintelligenties zich enkel richten tot andere superintelligenties. Ditzelfde kan gezegd worden voor de RUW-methode van Waser. Er kan niet vastgesteld worden dat een superintelligentie met ons gaat samenwerken om hun eigen doelstellingen te behalen. Wellicht hebben zij ons helemaal niet nodig en zijn ze beter af zonder ons. Tot slot zijn aporieën voor alle drie de theorieën een probleem. Situaties waar niet goed gedaan kan worden. Er zijn geen redenen te verzinnen waarom een KI hier wel tot een goed besluit kan komen.

Uit mijn scriptie is gebleken dat er twee grote uitdagingen zijn in het kader van waarden in KI: het formuleren en formaliseren. Ik heb aangetoond dat de theorieën van Bostrom, Yudkowsky en Waser niet geschikt zijn voor implementatie in de KI. De formulering van dit drietal levert problematische uitkomsten en de gevaren die zijn beschreven in hoofdstuk 2 worden hiermee niet voorkomen.

De auteurs die ik in mijn scriptie heb besproken lijken erg optimistisch te zijn als het gaat over de komst van superintelligentie. Daarnaast lijken zij ervan overtuigd dat wanneer superintelligentie bereikt wordt, de machines ook hetzelfde denken en redeneren als wij. Dit *an sich* is al een problematische aanname. Mochten we robots kunnen ontwerpen die algemene intelligentie bezitten, kunnen we nog niet aannemen dat deze gelijk is aan de menselijke intelligentie. Verwerkingsprocessen in de computer zijn nou eenmaal anders dan in onze hersenen. Ook de lichamelijkheid verschilt heel erg van die van de mens. Bepaalde lichamelijke waarden zullen niet gelden voor een robot. Een arm verliezen is voor een robot niet problematisch, die kan vervangen worden. Dit soort details hebben grote impact op de ethiek die mens en machine uitdragen.

Een begrip van de natuurlijke taal is naar mijn mening een ander punt dat nodig is voor implementatie van KI in de wereld. In films als *I, Robot* en *Elysium* zie je dat de robots altijd moeiteloos met mensen kunnen communiceren. Iets wat bij de Turing Test ook nodig is om als winnaar uit de bus te komen. Bij het 'Oekraïense jongetje' dat won bij de Loebner prijs kan bijna gezegd worden dat er is valsgespeeld. Juist door dommigheid was deze in staat de deelnemers te misleiden. Bij de robot Sophia lijkt het alsof zij de Engelse taal volledig begrijpt, maar ook zij zit er soms qua antwoorden volledig naast. Hoe gaan we met een robot communiceren als hij ons niet kan begrijpen wanneer we praten? Iedereen zou dan programmeertaal moeten beheersen. Dit is niet voor de hand liggend.

Tot slot is het twijfelachtig of het mogelijk is waarden te creëren zonder emoties. Waarden programmeren levert al een hoop problemen op, laat staan het formaliseren van emoties. Emoties zijn voor de mens beweegredenen om acties wel of juist niet te ondernemen. Dit is ook hoe normen en waarden ontstaan. De laatste jaren worden emoties veel meer geassocieerd met moraliteit in de filosofie. Dit is een onderwerp waar men ook aandacht aan moet besteden in de toekomst van de KI.

Ik hoop met deze scriptie het belang van menselijke waarden in KI aangetoond te hebben. De gevaren kunnen nog steeds catastrofaal zijn als we luisteren naar de besproken auteurs. Bostrom, Yudkowsky en Waser zijn daarmee niet geslaagd het mogelijke gevaar dat KI zich tegen de mens zou richten afdoende te ondervangen. Volgens velen is superintelligentie in de toekomst mogelijk. Het is daarom nodig dat de filosoof en de wetenschapper in discussie gaan over waarden binnen de KI, om zo tot een ethisch beginsel te komen dat wel toereikend is. Moraliteit voor KI is filosofie met een deadline. De komst van superintelligentie moet voor mogelijk gehouden worden. Het is daarom noodzakelijk te weten welke methode geschikt is bij het aanleren van waarden en vervolgens te weten welke waarden dat zijn.

Bibliografie

Bostrom, Nick and Yudkowsky, Elezier. “The Ethics of Artificial Intelligence.” in *The Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William M. Ramsey, 364-85. Cambridge: Cambridge University Press, 2014.

Bostrom, Nick. *Superintelligence.* Oxford: Oxford University Press, 2014.

Franklin, Stan. “History, Motivations, and Core Themes.” in *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 32-52. Cambridge: Cambridge University Press, 2014.

Kurzweil, Ray. *The Singularity is Near: When Humans Transcend Biology.* New York: Viking, 2005

Mul, Jos de. “Waarom we robots zo vrezen.” *Trouw,* September 19, 2015.

Nietzsche, Friedrich. *Genealogy of Morals*. Translated by Ian Johnston. Virginia: Richer Resources Publications, 2009.

Russel, Stuart. *Human Compatible.* New York: Viking, 2019.

Sternberg, Robert. “Intelligence.” in *The Cambridge Handbook of Thinking and Reasoning*, edited by Keith J. Holyoak and Robert G. Morrison, 751-73. New York: Cambridge University Press, 2005.

Tegmark, Max. *Life 3.0. Being Human in the Age of Artificial Intelligence*. Amsterdam: Maven Publishing, 2017.

Turing, Alan. “Computing Machinery and Intelligence.” *Mind* 59, no. 236 (October 1950): 433-60.

Waal, Frans de. *Primates and Philosophers: How Morality Evolved.* Princeton University Press: Princeton, 2006.

Waser, Mark. “Rational Universal Benevolence: Simpler, Safer, and Wiser than “Friendly AI”.” in *Artificial General Intelligence*, edited by R. Goebel, J. Siekmann, and W. Wahlster, 153-62. Heidelberg: Springer, 2011.

Yudkowsky, Eliezer. “Complex value systems in Friendly AI.” in *Artificial General Intelligence,* edited by R. Goebel, J. Siekmann, and W. Wahlster, 388-93. Heidelberg: Springer, 2011.

Yudkowsky, Eliezer. *Coherent Extrapolated Volition.* San Francisco: Singularity Institute, 2004.

Yudkowsky, Eliezer. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* San Francisco: Singularity Institute, 2001.