DANIEL DENNETTS ARTIFICIËLE INTELLIGENTIETHEORIE

Een zoektocht naar relevantie en vruchtbaarheid

Naam:	Gislane Dijkstra
Studentnummer:	411764
Scriptiebegeleider:	dr. T.K.A.M. De Mey
Scriptieadviseur:	Prof. dr. J. De Mul
Leerstoelgroep:	Theoretische Filosofie
Woordenaantal:	10607
Datum:	Juli 2018

Erasmus Universiteit Rotterdam,
Faculteit der Wijsbegeerte

"Any theory that makes progress is bound
to be initially counterintuitive."
– Daniel C. Dennett, *The Intentional Stance*

Inhoudsopgave

Inleiding

Het vakgebied van Artificiële Intelligentie (AI) is relatief nieuw. Vooralsnog blijven ontwikkelingen elkaar in rap tempo opvolgen. Een kritische blik op deze ontwikkelingen is noodzakelijk om eventuele tekortkomingen en risico's bijtijds te achterhalen en hier actie op te ondernemen. Doordat AI naar verwachting onderdeel zal worden van het dagelijks leven en van de consumptiemaatschappij, dient zij met enige voorzichtigheid behandeld te worden. Dit houdt in dat onder andere de "black box" van AI geopend zal moeten worden om transparantie te creëren voor het publiek en om tot een algemener en duidelijker begrip van het domein te komen. Tegelijkertijd is het van belang dat critici van AI serieus worden genomen, en dat multidisciplinaire theorieën binnen het vakgebied van cognitieve wetenschap, computerwetenschap en filosofie zorgvuldig geanalyseerd worden.

In de filosofie zijn nog weinig relevante theorieën te vinden die zich specifiek richten op Artificiële Intelligentie, vanwege het korte bestaan van deze discipline. Het is derhalve van belang om de bestaande theorieën te doorgronden om hun toegevoegde waarde uit te lichten en kenbaar te maken. In deze thesis zal ik mij specifiek richten op Daniel Dennett, een van de filosofen die zich al sinds de vroege jaren 80 bezighoudt met het analyseren van Artificiële Intelligentie. Gezien Dennetts verleden met AI en zijn werkzaamheden kan in ieder geval worden gesteld dat hij in de jaren 90 en begin 21ste eeuw aantoonbaar waardevolle bijdragen heeft geleverd. De vraag die resteert is of Dennetts theorie, gezien de hedendaagse ontwikkelingen, nog steeds een belangrijke rol speelt.

Het doel van deze thesis is dan ook te onderzoeken wat de verdiensten en tekortkomingen zijn van Dennetts theorie over AI. De onderzoeksvraag luidt daarom als volgt: In hoeverre zijn de analyses van Daniel Dennett betreffende Artificiële Intelligentie (nog) relevant met betrekking tot hedendaagse ontwikkelingen?

Het antwoord op deze vraag beslaat vier hoofdstukken, waarin ik nader toelichting geef op bijbehorende en relevante informatie met betrekking tot deze onderzoeksvraag. Hoofdstuk één richt zich op de achtergrond van AI, waardoor duidelijk wordt wat de ontwikkelingen zijn en hoe deze elkaar op logische wijze opgevolgd hebben. Dit is relevant omdat een accurate interpretatie vergt dat Dennetts theorie in zijn historische context wordt geplaatst. Hoofdstuk twee gaat dieper in op Dennetts analyse van AI door de jaren heen. Hier wordt duidelijk dat Dennett zijn theorie door de jaren heen heeft verfijnd en geactualiseerd. In hoofdstuk drie worden de hedendaagse ontwikkelingen op het gebied van AI geanalyseerd, inclusief een voorzichtig toekomstperspectief. Hoofdstuk vier beschrijft Dennetts theorie in het licht van de hedendaagse ontwikkelingen. Uit deze thesis blijkt dat de theorie van Dennett nog altijd vruchtbare componenten bevat die aantoonbaar waardevol zijn bij de verdere ontwikkeling en benadering van AI.

1. De Geschiedenis van Artificiële Intelligentie

Artificiële Intelligentie zoals wij die nu kennen, kan het best worden begrepen in samenhang met haar opkomst en ontwikkeling. Dit hoofdstuk zal dan ook bestaan uit een korte uiteenzetting van de geschiedenis van AI, om zo tot een beter begrip te kunnen komen van de huidige ontwikkelingen.[1]

1.1 De Beginfase

Het concept van AI vindt zijn grondslag begin jaren 50. De tegenwoordig gebruikte benaming volgt pas een half decennium later. In 1950 publiceren zowel Claude Shannon als Alan Turing – onafhankelijk van elkaar – een artikel waarin de limieten van computers op de proef worden gesteld. Shannon richt zich in *Programming a Computer for playing Chess* op de mogelijkheid dat een computer op zelfstandige en gevorderde wijze kan schaken tegen een mens. Turing vraagt zich in *Computing Machinery and Intelligence* af of computers over een denkvermogen beschikken. Hij bedenkt een gedachte-experiment waarbij een subject via een computer een gesprek voert met zowel een mens als een computer. Het subject moet bepalen welke van de twee de computer is. Zowel Shannon als Turing, beiden voorloper op hun tijd en beschikbare technologieën, voorspellen dat de potentie van een computer om niet meer onderscheiden te kunnen worden van het menselijk intellect, ooit gerealiseerd zal worden. Deze twee artikelen, dat van Turing in het bijzonder, veroorzaken een domino-effect in het onderzoek naar en de ontwikkeling van AI.

De benaming "Artificiële Intelligentie" is voor het eerst gebruikt in 1955 door John McCarthy, gemotiveerd door de gedachte dat "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (2006, p. 12). McCarthy had hierbij een machine in gedachten die in staat is om voornamelijk logische taken te verrichten, zoals berekeningen uitvoeren. De Dartmouth Conferentie, waar McCarthy de definitie van AI heeft geïntroduceerd, wordt veelal beschouwd als de geboorte van AI (Crevier, 1993). Nadat AI in 1956 officieel tot vakgebied wordt benoemd, volgt een reeks belangrijke ontwikkelingen binnen de discipline.

De volgende jaren laten een uitzonderlijk optimisme over de ontwikkeling van AI zien. Meerdere wetenschappers en filosofen zijn ervan overtuigd dat AI binnen enkele decennia haar piek zal bereiken, wat inhoudt dat een machine dan zodanig intelligent zal zijn dat ze op hetzelfde niveau presteert als een mens. Grote instanties investeren in onderzoek naar AI. Een fors deel wordt gebruikt voor de ontwikkeling van een linguïstiek vermogen, om communicatie met de mens mogelijk te maken.

[1] In dit hoofdstuk zal ik uitsluitend filosofisch relevante ontwikkelingen behandelen. De technische kant van de ontwikkeling van AI is niet van belang om mijn onderzoeksvraag te kunnen beantwoorden.

Zo volgt eind jaren 60 de eerste chatbot, ELIZA.[2] Binnen de robotica werd in 1973 de eerste mijlpaal bereikt, wanneer WABOT-1[3] werd gebouwd. Dit was de eerste androïde die kon lopen en simpele menselijke handelingen kon verrichten (Kato, z.d.). Ook kon deze robot door middel van een kunstmatige mond en linguïstiek vermogen een gesprek voeren in het Japans. Door deze ontwikkeling werd het abstracte concept van AI verenigd met een concrete uitvoering.

Het optimisme over AI was echter ongegrond, en vanaf de jaren 70 strandt het onderzoek. Het creëren van AI bleek ernstig onderschat, terwijl de lat enorm hoog was gelegd. Het uitblijven van veelbelovende resultaten zorgde ervoor dat de financiële steun werd stopgezet (Crevier, 1993), wat resulteerde in een vicieuze cirkel. Tegelijkertijd stonden er verschillende academici op die sceptisch waren over AI, waaronder Hubert Dreyfus en John Searle.

1.2 Critici van AI

Hubert Dreyfus is een van de meest radicale critici van AI. In 1979 publiceert hij *What Computers Can't Do,* waarin hij het onderzoek naar AI openlijk bekritiseert. Zo stelt hij over de stagnatie van het onderzoek naar AI dat "…the general failure to fulfill earlier predictions suggest the boundary may be near" (p. 139), en dat "current stagnation should be grounds for pessimism" (p. 139). Dreyfus' belangrijkste argument tegen de mogelijkheid van een intelligente machine is, wat hij noemt, het onderscheid tussen "weten-dat" en "weten-hoe" (1986).[4] Wanneer wij weten-dat, grijpen wij aan op ons bewustzijn en plannen wij de stappen die nodig zijn om een probleem op te lossen. Dit vereist zorgvuldige analyse die wij door middel van logica en linguïstisch vermogen voltrekken. Weten-hoe, anderzijds, vereist geen dergelijk rationeel denkvermogen. De handelingen die door middel van weten-hoe worden verricht, zijn geworteld in ons onderbewustzijn en worden geautomatiseerd uitgevoerd. Hierdoor lijkt deze manier van actie ondernemen eerder op intuïtie, mede doordat deze handelingen grotendeels beïnvloed worden door onze persoonlijke en unieke achtergrond. Dreyfus beargumenteert dan ook dat alhoewel weten-dat mogelijk is voor een intelligente machine, weten-hoe altijd menseigen zal zijn. Het gevolg is volgens Dreyfus dat AI onontkoombaar zal falen (1992).

In 1980 levert John Searle in *Mind, brains, and programs* eveneens een waardevolle bijdrage aan onze verwachtingen van AI. Hij stelt zich voor dat hij zich in een opgesloten kamer bevindt met louter een reeks Chinese tekens, waar hij geen verstand noch begrip van heeft. Hierna volgt een tweede set, waarop in de moedertaal staat uitgelegd hoe met de Chinese tekens omgegaan moet worden om correcte

[2] ELIZA was oorspronkelijk gecreëerd door Joseph Weizenbaum als parodie. Weizenbaum had als doel om aan te tonen dat de communicatie tussen mens en machine oppervlakkig en ongeloofwaardig is, maar het tegendeel bleek waar te zijn.

[3] WABOT is een samentrekking van Waseda Robot. De Waseda Universiteit in Japan heeft deze robot ontwikkeld.

[4] Deze twee definities zijn in overeenstemming met en kunnen verlengd worden naar het onderscheid tussen feitenkennis en vaardighedenkennis.

zinnen te kunnen formuleren. Op basis van de set Chinese tekens en bijbehorende instructies kunnen nu vragen worden beantwoord, waarbij de tekens in goede volgorde worden gelegd. Searle benadrukt dat er hier echter absoluut geen sprake is van enige vorm van begrip; er wordt louter een computationele taak uitgevoerd op basis van formeel gespecificeerde elementen. Voor de discipline van AI betekent dit, volgens Searle, dat een kunstmatige machine dus geen inherent begrip hoeft te hebben van de wereld. Ze voert slechts computationele en formele taken uit en formuleert op basis hiervan een antwoord, zonder dat het antwoord zelf betekenis heeft voor de machine. Voor Searle houdt deze conclusie in dat de Turing Test dus niet op adequate wijze aan kan tonen dat een computer over intelligentie kan beschikken.

Dreyfus en Searle vormen een belangrijk kritisch potentieel voor de verdere ontwikkeling van AI. Zoals in hoofdstuk 4 duidelijk zal worden is de kritiek van Dreyfus en Searle verwerkt in de verdere ontwikkeling van AI, waardoor problemen worden overwonnen die in de eerste 30 jaar van AI een belangrijke (en belemmerende) rol speelden.

1.3 Connectionisme

In de late jaren 80 rukt een nieuwe stroming op die relevant is voor onder andere de benadering van AI: het connectionisme. Deze stroming tracht door middel van kunstmatige neurale netwerken de intellectuele capaciteiten van ons brein uit te leggen. Wat deze stroming uniek en waardevol maakt, is dat ze op een nieuwe manier cognitieve processen verklaart. Hierbij vormt ze een alternatief voor de klassieke opvatting dat cognitie een symbolisch taalproces is, vergelijkbaar met dat van een digitale computer (Garson, 2016).

Het connectionisme veronderstelt dat het brein bestaat uit een netwerk van neuronen en de verbindende synapsen. Een neuron is hierbij een stukje informatie dat door de synapsen wordt vervoerd naar een ander neuron. Hierdoor wordt de menselijke cognitie gevormd en in werking gesteld. Dit betekent dat externe input processen activeert in ons brein, welke weer worden omgezet in fysieke output. Op deze manier is het neurale netwerk een alomvattend systeem: het verklaart hoe wij externe stimuli ontvangen, verwerken en hier actie op ondernemen.

Als het connectionisme waar is, dan is dit veelbelovend voor de ontwikkeling van AI. Er moeten dan eenheden worden opgesteld die lijken op neuronen, die door middel van kunstmatige synapsen kunnen communiceren. Het eerste prototype hiervan bestond uit twee lagen. De bovenste laag ontvangt externe input, en deze input wordt weer doorgegeven aan de onderste laag, waar hij verwerkt wordt. Deze simpele vorm van informatieverwerking werd een perceptron genoemd en was voornamelijk baanbrekend door het snelle karakter: perceptrons werken *real-time*, waardoor een AI net zo snel informatie kan verwerken en hierop kan reageren als een mens. De echte doorbraak volgde echter pas bij de ontwikkeling van een drielagig neuraal systeem, waardoor patronen accurater herkend

worden en deze zelfs gecategoriseerd kunnen worden (Dooremalen, De Regt, & Schouten, 2011). Dit heeft als voordeel dat er naast het verwerken en karakteriseren van een input ook ruimte is voor uitzonderingen. Zo kan bijvoorbeeld een tijger, normaliter een oranje met zwart zoogdier, ook nog een tijger worden genoemd wanneer hij albino is en dus niet oranje met zwart (Garson, 2016, par. 4).

Er werd echter ook kritiek geleverd op het connectionisme. Zo wordt gesteld dat alhoewel het connectionisme neuronen en netwerken als uitvalsbasis heeft, er geen rekening wordt gehouden met de aparte functies van deze neuronen. Ook worden hormonen en neurotransmitters volledig buiten spel gelaten, terwijl deze wel degelijk van invloed zijn op het functioneren van onze cognitie (Garson, 2016). Het connectionisme beperkt zich hier dus onterecht tot de basale functies van het menselijk brein, wat geen accuraat beeld schetst van de werkelijkheid. Het belangrijkste kritiekpunt tegen connectionisme is dat neurale netwerken zich niet aan formele regels lijken te binden, wat een logisch systeem onmogelijk maakt. Dit is problematisch wanneer kunstmatige neurale netwerken gebruikt worden binnen AI. Als laatste wordt niet duidelijk hoe het drielagige neurale netwerk bij AI ingezet kan worden als *backpropagation*[5] (Garson, 2016). Dit laatste probleem lijkt echter nog altijd relevant, zoals bij *machine learning* en het latere *deep learning*. Hier weid ik in hoofdstuk 3 verder over uit.

Gezien de geschiedenis van AI en de verdere ontwikkeling is het connectionisme de eerste serieuze poging om een machine zo menselijk mogelijk te maken. Het connectionisme is op dit gebied verder dan andere stromingen, door een oplossing voor te stellen voor het *real-time* kunnen verwerken van informatie. De vraag blijft of de menselijke cognitie wel te definiëren valt als een set met logische regels, zoals critici tegenspreken.

[5] Deze term heeft geen Nederlandse vertaling. Ruwweg gesteld is *backpropagation* een algoritme in een AI-machine dat ervoor zorgt dat bij het identificeren van een object een schatting wordt gemaakt van de kansen dat het object iets wel of niet is. Er wordt een gewicht gegeven aan elke mogelijkheid, waarbij de ene mogelijkheid waarschijnlijker zal zijn dan de andere. Dit is van belang voor het kunnen maken van accurate voorspellingen.

2. Daniel Dennetts Artificiële Intelligentie theorie

Daniel Clement Dennett is een Amerikaans filosoof en cognitiewetenschapper die zich vanaf eind jaren 70 bezighoudt met het bewustzijn, robotica en artificiële intelligentie. Dennett is een zelfbenoemde aanhanger van het verificatiebeginsel, dat stelt dat kennis cognitief waardevol is dan en slechts dan als het herleidbaar is tot zintuigelijke waarneming (Creath, 2017, par. 4.1), of wanneer het tautologisch verifieerbaar is. Tegelijkertijd sluiten de ideeën van Dennett, tenminste gedeeltelijk, aan bij die van het functionalisme. Deze stroming stelt dat mentale toestanden zijn opgebouwd uit een verzameling functionele relaties, die in relatie staan met andere mentale toestanden (Levin, 2017). Het wordt duidelijk dat Dennett ideeën heeft overgenomen van deze twee theorieën wanneer hij spreekt over artificiële intelligentie. Dit hoofdstuk zal dan ook bestaan uit een compacte uiteenzetting van zijn ideeën hierover, te beginnen vanaf eind jaren 70. Gezien de omvang van deze thesis komen slechts de hoofdlijnen van zijn theorie die bijdraagt aan het vakgebied van AI, aan de orde.

2.1 De Eerste Voorspellingen

Zoals de titel al suggereert, maakt Dennett in *Artificial Intelligence as Philosophy and as Psychology* (1978) – onderdeel van zijn essaybundel *Brainstorms* – onderscheid tussen filosofie van AI en psychologie van AI. Het belangrijkste verschil is volgens hem dat filosofie een abstracte en epistemologische aanpak heeft, terwijl psychologie op inductieve wijze te werk gaat. Dit wil zeggen dat er in een specifiek systeem wordt gekeken naar de noodzakelijke kenmerken, die vervolgens op universele wijze toegepast worden (p. 112). De beste aanpak ligt volgens Dennett in het midden: gestelde vragen over AI dienen abstract te blijven, maar modellen moeten specifiek en expliciet zijn (p. 113). Gezien het psychologische aspect van AI is volgens Dennett het ultieme doel van AI “to provide a hierarchical breakdown of parts in the computer that will mirror or be isomorphic to some hard-to-discover hierarchical breakdown of brain-event parts” (p. 114). Dennett merkt hier op dat de eerste stap naar een functionerende AI is om af te bakenen welke aspecten van het menselijk brein gesimuleerd dienen te worden door een AI. Maar, stelt hij, hier is op dit moment (1978) nog geen sprake van. De bestaande modellen wijken in elk opzicht af van de menselijke natuur. Dennett vermoedt dat dit in de toekomst anders zal zijn. Zijn eerste voorspelling luidt dan ook dat er op een dag modellen worden gemaakt op neuraal niveau[6] (p. 114).

In *Where am I?*, eveneens onderdeel van *Brainstorms* (1978), onderzoekt Dennett de mogelijkheid dat het brein (en dus het bewustzijn) los van het lichaam kan functioneren. Hij maakt hierbij gebruik van een gedachte-experiment waarbij hij een breinoperatie krijgt die ervoor zorgt dat

[6] Ik wil extra benadrukken dat deze voorspelling door Dennett werd gedaan nog voor de opkomst van het connectionisme.

zijn brein, met behoud van alle cognitieve functies, in een vat staat opgeslagen. De conclusie die Dennett trekt is dat hoewel hij weet dat het brein in het vat van hem is, dit nagenoeg onmogelijk is om te ervaren. Hij kan zich simpelweg niet voorstellen dat zijn waarneming en bewustzijn buiten de rest van zijn lichaam plaatsvinden. Dennett becommentarieert zijn eigen gedachte-experiment later in *The Mind's I* (1981), en concludeert het volgende:

> It is not just that in order for a computer to come close to matching a human brain in speed of handling millions of channels of parallel input and output it would have to have a fundamental structure entirely unlike that of existing computers. Even if we had such a brainlike computer, its sheer size and complexity would make the prospect of independent synchronic behaviour virtually impossible. Without the synchronized and identical processing in both systems, an essential feature of the story would have to be abandoned. (p. 230)

Als Dennetts stelling waar is, heeft dit consequenties voor AI. Dennett is van mening dat het niet mogelijk is om het menselijk brein over te zetten in een kunstmatige machine waarbij de machine te allen tijde parallel blijft lopen aan het menselijk brein. Er zal dus een andere manier gezocht moeten worden om AI mogelijk te maken, in plaats van een methode waarbij het brein nauw in verbinding staat met een computer.

Dennett ziet echter nog een problematisch fenomeen op het gebied van AI. Deze kwestie is niet gerelateerd aan de informatieoverdracht van mens naar machine, maar richt zich op de verwerking van externe informatie door een machine. Dennett noemt dit het frame-probleem van AI (zie Dennett, 1984). Dennett vraagt zich af hoe een artificieel systeem onderscheid kan maken tussen relevante en irrelevante input, en stelt de epistemologische vraag wat überhaupt bepaalt of iets relevante of irrelevante informatie is. Zoals hij in *Cognitive Wheels* (1984) demonstreert, zijn de gevolgen van het frame-probleem immens groot voor de ontwikkeling van AI. Zo verliest een artificieel systeem niet alleen bijzonder veel van haar *real-time* verwerkingssnelheid door alle informatie te filteren, maar kan het foutief filteren van informatie ook desastreuze gevolgen hebben.

In de eerste werken van Dennett staan dus voornamelijk de vorm en uitvoering van AI centraal. Dennett is van mening dat het vakgebied van AI zich door zowel psychologie als filosofie moet laten leiden. De uitvoering van een kunstmatig systeem is echter nog problematisch in zijn optiek, gezien het onvermogen om een kunstmatig brein te simuleren. Zijn voorspelling dat neurale modellen hier een uitkomst kunnen bieden, wordt bevestigd door het connectionisme, waar Dennett dan ook op inhaakt. Het frame-probleem van AI blijft echter onopgelost.

2.2 Een Hybride Theorie

In zijn toenmalige magnum opus *Consciousness Explained* (1991) introduceert Dennett een geheel nieuwe theorie over het bestaan en werking van het bewustzijn: de *multiple drafts* theorie. Dennett tracht ons hiermee van een alternatief te voorzien voor het idee van een Cartesiaans theater (p. 17), dat hij beschrijft als “de plek waar alles samenvalt en het bewustzijn gebeurt” (p. 39). Dit – volgens Dennett – foute idee is geworteld in de manier waarop de mens over bewustzijn denkt. De *multiple drafts* theorie biedt hier een uitkomst, en veronderstelt dat bewustzijn niet op één plek en op één moment plaatsvindt, maar dat sensorische informatie parallel ons brein binnenkomt en ook parallel verwerkt wordt. Het brein wordt hier beschouwd als informatieverwerkingsmechanisme. In de praktijk houdt dit in dat nieuwe informatie als constant proces wordt toegevoegd aan ons wereldbeeld en hierbij onze opvattingen en handelingen beïnvloedt. Er is volgens Dennett dus geen reden om aan te nemen dat in het brein een homunculus zit die ons bewustzijn bepaalt en ons handelen aanstuurt. Het Cartesiaans dualisme wordt hier dus nietig verklaard door Dennett, en in plaats daarvan wordt ruimte gemaakt voor een bewustzijnstheorie die aansluit bij het fysicalisme.

Wederom in *Consciousness Explained* verlengt Dennett zijn bevindingen naar een belangrijk verschil tussen mens en computer. Dit verschil was voor critici tevens grond van hun argument tegen het kunnen bestaan van AI. Dennett legt uit dat het brein een parallelle architectuur heeft, dat wil zeggen dat neurale processen gelijktijdig gebeuren. Een computer daarentegen heeft een seriële architectuur, waarbij processen na elkaar gebeuren, en dus niet gelijktijdig (p. 215). Dennett weerlegt de kritiek tegen AI vrijwel direct door te stellen dat “een virtuele machine een tijdelijke set van uiterst gestructureerde regelmatigheden [is], opgelegd op de onderliggende hardware door een programma” (p. 216). Dit houdt in dat er eindeloos veel mogelijkheden zijn waarop de beschikbare informatie en instructies gecombineerd kunnen worden, wat resulteert in een enorme database waar de machine uit kan halen hoe ze moet reageren. Het is irrelevant of een computer parallel of serieel informatie verwerkt, gezien het resultaat. De volgorde van verwerking is louter een middel tot het doel, en het doel is het enige dat van belang is. Dennett maakt hierna een belangrijke observatie in zijn poging om mens en machine verder te vergelijken:

> “Conscious human minds are more-or-less serial virtual machines implemented – inefficiently – on the parallel hardware that evolution has provided for us.” (p. 218)

Het wordt hier duidelijk dat Dennett geen onoverkoombare verschillen ziet tussen mens en machine. Integendeel, Dennett is ervan overtuigd dat deze analogie helpt om het menselijk bewustzijn te begrijpen. Dit wordt niet verder behandeld in deze thesis, maar het is desalniettemin van belang om op te merken dat Dennetts benadering van AI in grote mate overeenkomt met zijn benadering van menselijk bewustzijn.

2.3 AI als Evolutionair Product

Dennett weidt verder uit over de minimale verschillen tussen mens en AI in *Darwin's Dangerous Idea* (1995). Ditmaal beschrijft hij zijn theorie vanuit evolutionair oogpunt. Het standpunt dat hij aanneemt, heeft als hoofdgedachte dat de mens zelf is gemaakt uit robots, of wat hij ook wel noemt "een verzameling van triljoenen macromoleculaire machines" (p. 206). Dennett tracht hier niet te verdedigen dat een mens zelf een robot *is*, maar slechts dat een mens afstamt van een robot, en dus uit robotische onderdelen bestaat. De conclusie waar Dennett hier naartoe werkt, is dat als de mens bestaat uit robotische deeltjes, en de mens heeft een bewustzijn, dan kunnen robots zelf ook een bewustzijn hebben. Dennett merkt op – in aanvulling hierop – dat het evolutionair proces algoritmisch is, dat wil zeggen dat het naar een bepaald resultaat toewerkt (p. 308). Ook de mens is dus geproduceerd door een algoritme in de natuur.[7]

Dennett legt verder uit waarom AI een evolutionaire aanpak vereist. Net zoals wij bij de natuur doen, moeten wij retrospectief kijken naar een product. We weten bijvoorbeeld niet precies waarom de mens de vorm heeft die hij heeft, we kunnen nu eenmaal niet terug in de tijd om te kijken wat er is gebeurd. Wat we echter wel kunnen doen, is proberen terug te redeneren wat hieraan vooraf is gegaan. Dit wordt ook wel *reverse engineering* genoemd. Dit is volgens Dennett mogelijk doordat de natuur mechanisch werkt, en verondersteld kan worden dat evolutionaire ontwikkelingen zijn gebeurd met een reden. Hetzelfde moeten wij volgens Dennett doen bij het ontwikkelen van een AI-machine. We hoeven niet precies vooraf te weten waar we mee bezig zijn en welke algoritmen welke effecten hebben, we moeten gewoon de sprong wagen en zien wat eruit komt. De uitkomsten kunnen dan weer geanalyseerd en teruggeredeneerd worden. Zo kunnen we volgens Dennett pas echt de werking van een artificieel systeem begrijpen (p. 212).

2.4 Het Cog Project

Tot 2003 heeft Dennett deelgenomen aan het Cog project, onderdeel van de Humanoïde Robotica Groep van het Massachusetts Instituut voor Technologie (MIT). De motivatie achter het Cog project was om een robot te bouwen die op natuurlijke wijze kon communiceren met mensen. Een van de doelen hiervan was dat de robot – net als kinderen – nieuwe dingen zou kunnen leren, onthouden en later ook toepassen op sociaal gebied. Voor de discipline van AI was dit bijzonder relevant omdat het ontwikkelen van een dergelijke robot revolutionair zou zijn binnen het vakgebied, gezien het – toentertijd - onvermogen van een AI-machine om dergelijke taken te kunnen verrichten. Deze nabootsing bleek echter op bepaalde vlakken gecompliceerder dan verwacht, met als gevolg dat in 2003 de stekker uit het project werd

[7] Dit houdt overigens niet in dat er een specifiek algoritme is voor het creëren van de mens. Dennett heeft het hier over een simpel algoritme – natuurlijke selectie – dat variatie, selectie en reproductie omvat.

getrokken. Desalniettemin heeft Dennett waardevolle contributies geleverd die niet alleen toegesneden zijn op Cog, maar die ook op algemeen niveau toegepast kunnen worden.

Allereerst legt Dennett in *Consciousness in Human and Robot Minds* (1997) uit hoe Cog, en dus potentieel ook andere robots, een bewustzijn kan hebben. Dennett benadrukt dat het noodzakelijk is dat een machine sensorische informatie kan ontvangen. Dit impliceert dat een AI-machine, om een bewustzijn te kunnen hebben, materieel moet zijn en niet alleen als software kan bestaan (p. 8). Dit is niet alleen nodig om informatie te kunnen verwerken, maar bijvoorbeeld ook om een AI van hand-oogcoördinatie te kunnen voorzien (p. 5). In het verlengde hiervan moet de machine zichzelf motorisch vermogen aan kunnen leren, dat wil zeggen dat het zich handig kan voortbewegen en niet herhaaldelijk tegen objecten botst. Gezien het kinderlijke karakter van een dergelijke AI-robot is er in het begin dus constante aanwezigheid vereist ter bescherming. Anderzijds is het onmogelijk dat een dergelijke intelligente robot als *tabula rasa* de wereld in wordt geholpen; er is altijd een set aan basiskarakteristieken nodig om ontwikkeling mogelijk te maken (p. 6). Dit zou bij mensen het DNA zijn. Dennett benadrukt dat voornamelijk linguïstiek vermogen essentieel is voor het creëren van een kunstmatig bewustzijn. Dit gecombineerd met ingebouwde 'persoonlijke' doelen en voorkeuren moet ervoor zorgen dat een AI-machine intentionaliteit kan vertonen. De hardware van een dusdanige machine moet ervoor zorgen dat informatie *real-time* verwerkt wordt en dat problemen direct kunnen worden opgelost. Dennett is optimistisch en voorspelt dat wanneer aan deze voorwaarden wordt voldaan zelfs sceptici toe zullen moeten geven dat er sprake is van een bewustzijn (pp. 7-8). In *Cog as a thought experiment* (1997) voegt Dennett nog toe dat Cog, of dus een AI-robot in het algemeen, een nieuwsgierig – Dennett noemt het ook wel 'Epistemisch hongerig' – karakter moet bezitten om te kunnen ontwikkelen tot het niveau van een volwaardige volwassene (p. 255). Zou hier geen sprake van zijn, dan zal de robot geen intentionaliteit vertonen; een voorwaarde voor bewustzijn.

2.5 Dennetts Hedendaagse Theorie

Dennett heeft zich sinds het Cog project weinig uitgelaten over AI, en zijn oeuvre vanaf 2003 op dit gebied bestaat voornamelijk uit interviews en krantenartikelen. Er wordt hier echter weinig nieuws verteld, en het merendeel van de standpunten die Dennett inneemt, zijn terug te leiden naar zijn voorgaande boeken.[8] In 2017 publiceert Dennett echter *From Bacteria to Bach and Back*, waarin hij zich weer volledig stort op het bewustzijn en AI.

In *From Bacteria to Bach and Back* contrasteert Dennett intelligent design (ID) met evolutie. ID is de gedachte dat er een intelligente schepper is die alle entiteiten van het heelal heeft gecreëerd. Er

[8] Dennett houdt zich in deze jaren voornamelijk bezig met een filosofische analyse van religie, en publiceert hier meerdere boeken en artikelen over. Hier weid ik niet verder over uit, om de reden dat er geen relevante informatie betreffende AI in staat.

is hier sprake van een bepaalde doelmatigheid, en zodoende spreekt men van een *top-down* benadering. Evolutie daarentegen is richtingloos en doelloos, en brengt slechts wijzigingen aan op datgeen wat zij reeds heeft gecreëerd (p. 109). In dit geval is dus sprake van een *bottom-up* benadering. Het punt dat Dennett hier maakt is dat alhoewel de evolutie geen AI of computers heeft voortgebracht, dit niet betekent dat er ID aan te pas is gekomen. Evolutie heeft ons voorzien van hersenen die in staat zijn om AI en computers te ontwerpen. De mens is als het ware dus zelf een ID'er. Het is volgens Dennett zinloos om *reverse engineering* toe te passen op evolutie, gezien het complexe karakter ervan. Bij het ontwikkelen van een softwareprogramma kan echter hetzelfde gesteld worden, zoals Dennett uitlegt:

> If we make the effort to decipher spaghetti code, we can usually note which unlikely possibilities *never occurred to* the designers in their myopic search for the best solution to the problems posed to them. *What were they thinking?* When we ask the same questions about Mother Nature, the answer is always the same thing: nothing. No thinking was involved, but nevertheless she muddled through …" (p. 116)

Dit leidt op het eerste gezicht tot een paradox: door middel van *bottom-up* evolutie zijn wij in staat om *top-down* te ontwerpen, maar door middel van *reverse engineering* valt in beide gevallen geen feilloze kennis te halen. Dennett maakt hier dus wederom gebruik van een evolutionaire benadering, maar past ze ditmaal toe op software. Zijn conclusie is dat het pad van evolutie naar mens, en vervolgens naar computers en software geen onderliggende structuur of logica heeft die wij kunnen begrijpen, maar desalniettemin is het een goed functionerend systeem. Dit lijkt in tegenspraak te zijn met zijn eerdere standpunt in *Darwin's Dangerous Idea* (1995), waarin hij verdedigt dat *reverse engineering* een waardevolle aanpak is om tot een beter begrip te komen van de processen die voorafgingen aan een bepaalde uitkomst. In Hoofdstuk 4 zal echter duidelijk worden dat dit juist aansluit bij zijn meest recente analyse.

Naast het onderscheid tussen *bottom-up* evolutie en *top-down* programmeren, merkt Dennett verschillen op tussen brein en computer. Het belangrijkste verschil voor hem is dat breinen leven, en computers niet. Dennett stelt dat informatie slechts een bron van energie nodig heeft om tot stand te komen, en erkent dus dat informatie mediumeutraal is (p. 205). Of dit neuronen zijn of elektriciteit is voor Dennett dus niet van belang. Wat volgens Dennett de relevantie van het verschil tussen levend en niet-levend dan wel is, heeft te maken met de energietoevoer. Computers en andere kunstmatige informatiesystemen zijn parasitair volgens Dennett, wat wil zeggen dat ze qua energietoevoer afhankelijk zijn van een gebruiker. Laadt de gebruiker het apparaat niet op of wordt het niet aangesloten op het stroomnet, dan werkt de machine niet. Breinen functioneren echter autonoom[9], en zijn tevens

[9] Er kan natuurlijk altijd gesteld worden dat mensen hun energie uit voeding halen, maar voor Dennetts argument is dit verder niet van belang.

samengesteld uit levende cellen die ook weer autonoom zijn (p. 207). Het ultieme verschil tussen mens en AI-machine is volgens Dennett dus dat mensen onafhankelijk van machines kunnen bestaan, maar andersom niet. Het wel autonoom maken van een AI-machine schiet het doel voorbij volgens Dennett, en maakt het creëren van een AI onnodig complex. Dennett lijkt hier ten opzichte van eerdere werken zijn standpunt dat mens en machine gelijk zijn, bij te stellen. Dit geldt echter alleen voor het materiële vlak, het immateriële vlak laat hij hier in het midden.

Dennett sluit zijn boek – en tevens zijn oeuvre tot nog toe – af met een paar kritische opmerkingen over de omgang met AI. Ten eerste is Dennett van mening dat het te allen tijde duidelijk moet zijn wanneer er gecommuniceerd wordt met een AI, en niet met een mens. Hij neemt hier een radicale houding aan, en stelt zelfs, indien dit niet verduidelijkt wordt, dat "their creators should go to jail for committing the crime of creating or using an artificial intelligence that impersonates a human being" (p. 501). Dennett pleit verder voor transparantie in de consumptiemaatschappij wat betreft de tekortkomingen en grenzen van een AI. Zo mag een kunstmatig intelligent apparaat nooit mooier gemaakt worden dan het is. Ook de risico's moeten worden belicht. Verder vindt Dennett dat degenen die AI gebruiken om mensen te beïnvloeden, trainingen moeten krijgen en aansprakelijk gehouden moeten worden voor eventuele uitkomsten. Deze behoedzame houding is voor Dennett essentieel om AI's de baas te blijven. Dit geldt voor hem overigens niet op het gebied van technologische singulariteit. Dit houdt in dat er in de toekomst een machinale superintelligentie zal ontstaan die iedere menselijke vorm van intelligentie overschrijdt. Dennett is hier sceptisch over. Voor hem dient de behoedzame houding dan ook slechts om wederzijdse afhankelijkheid tussen machine en mens goed te laten verlopen. Zoals Dennett zelf concludeert:

> And if the future follows the trajectory of our past – something that is partly in our control – our artificial intelligences will continue to be dependent on us even as we become more warily dependent on them. (p. 513)

2.6 Conclusie

Het doel van dit hoofdstuk was om de theorieën van Dennett betreffende AI vanaf eind jaren 70 tot nu uiteen te zetten. Zoals is gebleken, heeft Dennett door de jaren heen een uitgebreide theorie geformuleerd die mee lijkt te gaan met de daadwerkelijke ontwikkelingen. Dennett betreedt het vakgebied met een sceptische houding en stelt dat er meerdere problemen zijn die opgelost moeten worden, voordat een machine daadwerkelijk intelligent, dan wel bewust, kan worden genoemd. Later brengt Dennett zijn AI-theorie voornamelijk onder bij zijn bewustzijnstheorie, waardoor de twee niet meer los van elkaar te zien zijn. Dennett zelf begint brein en machine ook als nauw verwant aan elkaar te zien. Hij doet zelfs de radicale uitspraak dat mens en machine niet verschillen van elkaar. Hij beschouwt het bewustzijn echter wel als evolutionair product. Maar doordat de mens vervolgens weer

AI-machines heeft gecreëerd, stamt AI logischerwijs ook af van de evolutie. In zijn laatste boek sluit Dennett de cirkel van de opkomst van AI, het creëren van AI, de toepassing ervan en de uiteindelijke gevolgen. Wat zeker is, is dat Dennett de ontwikkelingen op nauwe voet heeft gevolgd. Zijn deelname aan het Cog project bevestigt dit expliciet. Het oeuvre van Dennett betreffende AI is rijkelijk gevuld en voltrekt zich over een periode van ruim dertig jaar. In dit tijdsbestek heeft hij een consistente theorie uiteengezet.

3. Artificiële Intelligentie in de Hedendaagse Tijd

Dit hoofdstuk bestaat uit een verheldering van relevante hedendaagse ontwikkelingen binnen het vakgebied van AI, aangevuld met enkele filosofisch relevante theorieën die hier betrekking op hebben. Het doel van dit hoofdstuk is om naar voren te brengen wat de *status quo* is van AI, en aan te tonen hoe deze logische opeenvolging voortvloeit uit de ontwikkelingen sinds de jaren 50.

3.1 Invloeden van het Connectionisme

Vanaf medio jaren 90 wordt afstand gedaan van artificiële neurale netwerken en de eerdergenoemde *backpropagation*. Onderzoekers zijn van mening dat deze neurale netwerken niet toepasbaar zijn op het creëren van een AI-machine, doordat het connectionisme neuronen en cognitie overgesimplificeerd weergeeft (Flusberg & McClelland, 2014, p. 3). Het afgelopen decennium is er echter een heropleving geweest van kunstmatige neurale netwerken. Dergelijke kunstmatige netwerken worden tot op de dag van vandaag breed ingezet bij het ontwikkelen en optimaliseren van AI-machines. Enkele voorbeelden van deze toepassingen zijn te vinden bij *machine learning* en *deep learning*.

3.1.1 Machine learning

In de jaren 90 bloeit een nieuwe methode op, die het mogelijk maakt dat een kunstmatig intelligent systeem zichzelf nieuwe informatie kan aanleren: *machine learning*. Het doel van *machine learning* is om alle input van de omgeving te gebruiken en hiervan te leren, zonder dat de machine specifiek geprogrammeerd is om dit te doen (Bennett & Hauser, 2013). Het grote voordeel van deze methode is dat veel minder code handmatig geprogrammeerd hoeft te worden om een kunstmatige machine nieuwe informatie te laten verwerken en gebruiken. *Machine learning* was de eerste stap richting een onafhankelijk functionerende machine. Bij *machine learning* wordt gebruikt gemaakt van algoritmen, dat wil zeggen een reeks instructies die een machine vanaf het startpunt tot een bepaald doel moeten leiden. Deze algoritmes kunnen op verscheidene manieren verwerkt worden door een machine, bijvoorbeeld door middel van een beslissingsboom of inductieve logica. Algoritmes op basis van kunstmatige neurale netwerken blijken echter het meest vruchtbaar, zoals bij *deep learning* duidelijk wordt. Bij *machine learning* wordt gebruik gemaakt van het – door het connectionisme voorgestelde – drielagige systeem, bestaande uit een input laag, een verborgen laag en een output laag. Een artificieel neuraal netwerk heeft hier een functie als netwerk van verschillende knopen (*nodes*), die allemaal met elkaar in verbinding staan. De verborgen lagen vormen de tussenschakel tussen de input- en output-laag (Rescorla, 2017, par. 4).

Kort gezegd stelt *machine learning* een kunstmatige machine dus in staat om informatie te verwerken, gebruiken, en om voorspellingen te kunnen doen op basis van deze data. Het is bij *machine*

learning nog wel noodzakelijk dat de AI-machine van tevoren geprogrammeerd wordt om deze informatie toe te kunnen passen, en dat er constant nieuwe data worden ingevoerd om de leercurve van het systeem te kunnen garanderen.

3.1.2 Deep learning

Deep learning bouwt voort op *machine learning*, en ze staat met haar bestaan van ongeveer een decennium nog in de kinderschoenen. Bij *deep learning* worden de basisprincipes van *machine learning* gebruikt, maar worden inductieve logica en andere vormen van algoritmisch leren, behalve artificiële neurale netwerken, verleden tijd. Dit brengt als groot voordeel met zich mee dat het niet langer nodig is, zoals bij *machine learning*, om alle losse stukjes data te labelen voordat een computer deze kan gebruiken.

Wanneer wordt gekeken naar de werking van *deep learning* wordt duidelijk dat het leidt tot een beter toekomstperspectief dan *machine learning*. Ruwweg gesteld bevat het proces van onbekende input tot betrouwbare identificatie bij *deep learning* zes fases:

1. De trainingsfase: Een machine wordt allereerst klaargestoomd om data te kunnen herkennen. Hierbij worden duizenden gespecificeerde stukken data ingevoerd, waardoor de machine deze data kan classificeren.

2. Input: Ongespecificeerde data wordt in de getrainde machine ingevoerd.

3. Eerste laag: De kunstmatige neuronen reageren op basale kenmerken, zoals vormen.

4. Hogere lagen: De kunstmatige neuronen reageren op complexere kenmerken, zoals structuur. De hogere lagen kunnen uit honderden segmenten bestaan.

5. Bovenste laag: De kunstmatige neuronen classificeren de meest complexe en abstracte eigenschappen. Dit kan bijvoorbeeld “man/vrouw” of “hond/wolf” zijn.

6. Output: Het netwerk voorspelt bij welke categorie de ingevoerde data worden ingedeeld, gebaseerd op de training die de machine heeft ondergaan (*backpropagation*).

Door middel van een beslisboom wordt een AI-machine telkens een stapje dichterbij de identificatie van ingevoerde informatie gebracht, wat leidt tot de uiteindelijke output. Neuronen vervullen hier dus de rol als tussenschakel van fases en als aftakkingen bij de beslisboom. Hoe meer een AI-machine deze stappen uitvoert, hoe accurater de output wordt. Dit komt doordat de data worden opgeslagen in de database van de machine, waardoor deze toegankelijk zijn voor de kunstmatige neuronen.

3.1.3 Beperkingen van artificiële neurale netwerken

De grote beperking van artificiële neurale netwerken is dat elk proces tussen input en output zich in een black box begeeft. Het is niet precies bekend hoe de artificiële neuronen werken en welk stukje informatie iedere neuron met zich meedraagt. Dit maakt het weer gecompliceerd om een kunstmatig intelligente machine te kunnen controleren. Een dergelijke machine leert zichzelf de meest efficiënte manier van verwerken, wat ertoe kan leiden dat de originele programmeertaal wordt overschreven. Wanneer de nieuwe processen dan ook niet meer terug te leiden zijn tot voor de mens begrijpbare taal, kan alleen nog het observeerbare gedrag van de AI-machine worden gebruikt om de werking in kaart te brengen. Dit is vanzelfsprekend niet ideaal. Er is dan ook behoefte aan het openen van de black box, voor zowel het efficiënt kunnen blijven ontwikkelen van algoritmes als het in de hand kunnen houden van een kunstmatig intelligente machine.

3.2 Overige Ontwikkelingen

Naast de vruchtbare methode van artificiële neurale netwerken zijn er andere methodes om een kunstmatig intelligente machine te programmeren en te laten functioneren. De meest – voor deze thesis – relevante theorieën worden hierna kort belicht.

3.2.1 Evolutionaire algoritmen

Evolutionaire algoritmen worden toegepast binnen het vakgebied van AI om optimalisatie- en zoekproblemen op te lossen. Dit houdt in dat algoritmen van deze soort gemaakt worden om tot kennis te komen over de manier waarop bepaalde natuurlijke processen zich hebben ontwikkeld en verbeterd in de loop der tijd. Op deze manier wordt onderzoek gedaan naar de efficiëntie van de natuur, en het tracht de onderliggende processen bloot te leggen. Een voorbeeld hiervan is het onderzoeken van de jaagtechniek van wolvenroedels, die door middel van evolutionaire algoritmen is verklaard en begrepen. Het algoritme dat in de AI-machine wordt geïmplementeerd wordt hier dus gebruikt als middel om *reverse engineering* mogelijk te maken.

3.2.2 Inductieve logica

Alhoewel inductieve logica gebruikt kan worden bij het programmeren van AI, blijkt dat dit in de praktijk nauwelijks leidt tot de gewenste resultaten. Gezien het feit dat een logische redenering binnen een bepaald systeem ontstaat, valt de validiteit van het gehele systeem weg indien een propositie binnen het systeem onwaar blijkt. Desalniettemin wordt de programmeertaal van inductieve logica veelal

gebruikt bij *natural language processing*[10]. Deze stroming binnen AI houdt zich bezig met het onderzoek naar de talige interactie tussen mens en computer, en onderzoekt hoe computers die hierop zijn toegespitst omgaan met taal en de opslag en verwerking hiervan. Inductieve logica werd hier gebruikt om formele regels op te stellen betreffende linguïstiek, welke de intelligente machine toepast. Dit bleek echter een zeer complexe taak, die dan ook grotendeels is vervangen door een combinatie van *machine learning* en statistische methoden (Mooney, 1997, p. 18).

3.3 Filosofische Perspectieven

Alhoewel filosofie geen actieve rol speelt binnen de ontwikkeling van AI, is juist de meta-positie die zij inneemt van belang. Het is dan ook niet ongegrond dat Nicholas Fearn (2007) opmerkt dat filosofie "twee hoofdrollen heeft in de huidige literatuur: om te bepalen of zulke machines bewust zijn, en om te voorspellen of zulke machines mogelijk zijn" (p. 55). Er zijn echter ook nog ethische implicaties die meegenomen dienen te worden in het filosofische debat betreffende AI.

3.3.1 Bewustzijn bij AI

De vraag wat de definitie van bewustzijn is, is evenmin beantwoord als de vraag of bewustzijn bij AI mogelijk is. In beide gevallen is tot op heden geen consensus bereikt, maar het is niet ondenkbaar dat een antwoord op de eerste vraag automatisch verheldering oplevert bij de tweede vraag. Er zijn drie populaire benaderingen die het bewustzijn trachten te verklaren, en welke allemaal een ander antwoord geven op de vraag of bewustzijn bij AI mogelijk is.

De eerste benadering van bewustzijn is materialistisch van aard. Deze benadering heeft als uitgangspunt dat er uit de samenwerking tussen onbewuste componenten (neuronen, microchips) een bewustzijn gevormd kan worden. Deze gedachte is veelal populair bij cognitieve wetenschappers, en wordt ook wel emergentie genoemd. Verder is er een bewustzijnstheorie die beïnvloed is door de kwantumfysica. Deze theorie heeft als hoofdgedachte dat het bewustzijn en de materiële wereld complementaire aspecten zijn van eenzelfde realiteit. Wanneer een persoon de fysieke wereld observeert, dan veroorzaakt de bewuste interactie van een persoon een waarneembare verandering. Bewustzijn is hier dus een voorwaarde voor fysiek waarneembare verandering.[11] De derde bewustzijnstheorie beschouwt het bewustzijn als reflexief. Het bewustzijn wordt hierbij als meta-cognitie gezien, en wordt uitgelegd als het besef (*awareness*) van het eigen gedachtegoed – ook wel zelfbewustzijn genoemd.

[10] Deze term kent geen accurate Nederlandse vertaling. Het meest overeenkomend zijn de termen 'computationele taalkunde' en 'computerlinguïstiek'.

[11] Het gedachte-experiment van Schrödingers kat berust op ditzelfde principe.

In het geval van de materiële bewustzijnstheorie is het goed voor te stellen dat AI een bewustzijn kan hebben. Het is immers denkbaar dat de verschillende componenten van een computer zodanig kunnen samenwerken dat er een bewustzijn ontstaat, net zoals de neuronen in het brein dit ook doen. Bij de bewustzijnstheorie van de kwantumfysica is dit eveneens mogelijk, aangezien het kwantumniveau slechts veronderstelt dat er een waarnemer aanwezig is. Er wordt in dit geval geen onderscheid gemaakt tussen een menselijke waarnemer en een artificiële waarnemer. De reflexieve bewustzijnstheorie is de meest testbare theorie van deze drie theorieën, en zal zich in de toekomst waar of onwaar bewijzen. Een dergelijke test kan bijvoorbeeld worden uitgevoerd door aan een kunstmatig intelligent systeem te vragen zichzelf te beschrijven. Indien het vervolgens een accurate beschrijving van zichzelf geeft – zonder dat het hiervoor getraind of geprogrammeerd is – kan wellicht gesproken worden van een bewustzijn.

Van deze drie theorieën is er echter geen een die het potentiële bewustzijn van AI afwijst. Dit lijkt een veelbelovend toekomstperspectief, maar op dit moment is het nog te vroeg om een eindoordeel te vellen. Ook zal naar alle waarschijnlijkheid, mocht AI zo ver zijn dat er ontwikkelingen zijn die vergelijkbaar zijn met het menselijk bewustzijn, niet voldaan worden aan alle criteria van de verschillende bewustzijnstheorieën. Het antwoord op de vraag of AI bewustzijn kan hebben is dus afhankelijk van de definitie van bewustzijn die gehanteerd wordt, maar wel is duidelijk dat op dit moment aan nog geen enkel criterium wordt voldaan.

3.3.2 Een ethische last dragen

Een van de meest vanzelfsprekende filosofische discussies over het ontwikkelen van AI is of de mensheid ethisch wel voorbereid is om AI-machines de wereld in te laten. Nu de toepassing van AI in het dagelijks leven geen ‘of’ meer is, maar een ‘wanneer’, buigt menig wijsgeer zich over de vraag wat de rol van een AI mag en moet zijn binnen de maatschappij. Er zijn meerdere risico’s die genomen worden bij het steeds humaner en intelligenter maken van een machine, en ieder risico dient serieus genomen te worden. Vanuit deze risico’s ontstaan de volgende belangrijke ethische vragen: Hoe te handelen indien AI voor ongewenste doelen wordt gebruikt? Kan (de maker van) een AI verantwoordelijk worden gehouden indien er onverwachte en ongewenste gevolgen zijn? Kan een superintelligente AI het eind van de mens betekenen? De antwoorden op deze vragen zijn, zoals normaliter het geval is bij ethische vraagstukken, niet eenduidig en vallen niet makkelijk te beantwoorden. Dit maakt het echter niet minder van belang om bij stil te staan, gezien de snelheid waarop het vakgebied van AI vooruitgang boekt.

3.4 Conclusie

Hedendaagse ontwikkelingen van AI richten zich voornamelijk op het vermogen van een artificieel systeem om zichzelf kennis aan te kunnen leren en zichzelf te verbeteren. Artificiële neurale netwerken blijken de meest vruchtbare manier om dit te kunnen verrichten. Het gevaar blijft echter dat de black box van artificiële neurale netwerken ongeopend blijft, en dat dit implicaties zal hebben voor de toekomst. Voor nu wordt er in de meeste gevallen gebruik gemaakt van verschillende methoden om *deep learning* te realiseren, waarbij zowel evolutionaire algoritmen en inductieve logica een rol spelen.

De toekomst van AI blijft vooralsnog onzeker. Er zijn meerdere bewustzijnstheorieën waar AI aan getoetst kan worden, welke allemaal verschillende criteria hanteren voor het bezitten van bewustzijn. Wat voor nu zeker is, is dat er volgens geen enkele theorie al sprake is van een bewustzijn bij AI. Het is vooralsnog belangrijk om hier niet op te wachten, en om zo snel mogelijk met een ethische aanpak voor AI te komen. De vragen hoe te handelen indien een kunstmatig intelligente machine zich afwijkend gedraagt, voor verkeerde doeleinden wordt gebruikt, en wie er dan verantwoordelijk is, dienen als leidraad te worden genomen. Op deze manier zijn we voorbereid op wat er nog kan komen.

4. Daniel Dennetts Theorie vs. Hedendaagse Theorie

In dit hoofdstuk volgt een vergelijking van Dennetts AI-theorie met de hedendaagse beschouwing en ontwikkelingen van AI en bewustzijn in het algemeen. Het doel is om aan te tonen dat Dennetts theorie compatibel is met de huidige stromingen. Hierbij wordt gebruik gemaakt van Dennetts theorie zoals in Hoofdstuk 2 uiteengezet, welke vervolgens gecontrasteerd wordt met de hedendaagse ontwikkelingen beschreven in Hoofdstuk 3. Hieruit zal blijken dat alhoewel Dennett op sommige punten afwijkt van hedendaagse optiek, er geen onoverkoombare verschillen zijn die zouden impliceren dat Dennetts theorie niet langer relevant zou zijn.

4.1 Van Voorspelling naar Werkelijkheid

Dennett begint zijn AI-theorie in 1978 met de stelling dat het vakgebied van AI zich voornamelijk moet laten beïnvloeden door psychologie en filosofie. AI is een interdisciplinair onderzoeksgebied, maar laat zich naast filosofie en psychologie tegenwoordig ook beïnvloeden door cognitieve wetenschap, neurowetenschap, computerwetenschap en linguïstiek. De voorspelling die Dennett in 1978 heeft gemaakt – dat neuronen een belangrijke rol gaan spelen bij het ontwikkelen van AI – blijkt dan ook zeer accuraat. Slechts enkele jaren na zijn voorspelling komt het connectionisme op, en tot op de dag van vandaag is de hoofdgedachte van deze stroming te vinden in ontwikkelingen als artificiële neurale netwerken die onder andere worden gebruikt bij *deep learning*.

Wanneer Dennett in 1981 concludeert dat het niet mogelijk is om een computer mensachtige, intelligente eigenschappen te laten vertonen, is hij in de veronderstelling dat het een voorwaarde is, dat computers parallel informatie kunnen verwerken. Tien jaar later lijkt hij deze aanname echter afgezwakt te hebben. Hij stelt dan dat zo lang een computer op seriële wijze even snel of sneller informatie kan verwerken als een mens dit op parallelle wijze doet, dit voor het einddoel geen verschil maakt. Een computer kan aldus intelligent zijn; er is immers nog altijd sprake van *real-time* verwerkingssnelheid. Deze bevinding gebruikt Dennett tevens als opstapje om zijn theorie op evolutionair niveau te verklaren. Zo beschrijft hij de evolutionaire drijfveer van ons bewustzijn als algoritme, die in deze zin deterministisch is. De mens is gedetermineerd om een bewustzijn te hebben, maar desalniettemin is dit bewustzijn niet onderhevig aan evolutionaire wetten die stellen dat elk macromoleculair deeltje geprogrammeerd is voor een specifiek doeleinde. Het bewustzijn is dus meer dan de som van de losse deeltjes, en als dit mogelijk is bij een mens, dan moet dit volgens Dennett ook mogelijk zijn bij een kunstmatig systeem. In slechts tien jaar tijd herziet Dennett zijn theorie dus grondig, en voegt hij zich bij het beperkte aantal academici dat overtuigd is van een toekomst met bewuste AI-machines. Deze bewustzijnstheorie van Dennett lijkt echter in grote mate overeen te komen met de materiële benadering van bewustzijn. Deze benadering stelt immers dat vele kleine componenten samen tot een nieuwe

eigenschap leiden, in dit geval bewustzijn. Deze benadering van bewustzijn heeft sinds eind jaren 70 aanzien binnen de cognitieve wetenschap en is sindsdien de dominante opvatting (McLelland, 2010). Dennett heeft dus ook in dit opzicht aantoonbaar inzicht in ontwikkelingen binnen het gebied van cognitiewetenschap. Dit kan tevens worden verlengd naar het vakgebied van AI, dat zich inspant om het tot op heden ongespecificeerde bewustzijn te simuleren.

Dennett neemt vooralsnog een fysicalistische houding aan ten opzichte van het bewustzijn, hij verklaart deze immers door een monistische stelling te poneren, namelijk dat elk aspect van het bewustzijn verklaard kan worden door lichamelijke processen. Dit doet hij in tegenstelling tot David Chalmers, die gebruik maakt van het zombie-argument om aan te tonen dat een louter fysicalistische houding ten opzichte van het bewustzijn incorrect is. Het standpunt dat hij hierbij wil verdedigen, is dat subjectieve ervaring, een van de criteria voor bewustzijn (Aleksander, 1995), niet fysicalistisch te verklaren is. Dit gedachte-experiment veronderstelt dat het mogelijk is om een nabootsing te maken van een bewust subject. Deze nabootsing – ofwel zombie – is identiek aan het bewuste subject, fysiologisch en fysiek. De zombie is eveneens gelijk aan het subject op functioneel en psychologisch niveau: “Hij zal de bomen buiten waarnemen, in de functionele zin, en de chocolade proeven, in de psychologische zin” (1995, p. 84). Er is echter een belangrijk verschil tussen zombie en subject: de zombie zal op geen enkele manier over een bewustzijn beschikken. Chalmers is hier niet geïnteresseerd in de vraag of het plausibel is dat dergelijke zombies kunnen bestaan, maar hij tracht te verdedigen dat het denkbeeld van een zombie van deze soort coherent is (p. 85). Dit doet hij door te stellen dat het een logische mogelijkheid is dat deze filosofische zombie kan bestaan, het is immers denkbaar, en dus is er geen reden om aan te nemen dat er sprake is van een bewustzijn. Voor Chalmers volgt hieruit dat het fysicalisme incorrect moet zijn; er is sprake van een entiteit die fysiek gelijk is aan een bewust subject, behalve dat de entiteit niet over een bewustzijn beschikt. Het fysicalisme is dus niet bevredigend, en dit is voor Chalmers reden om een dualistische houding aan te nemen.

Dennett zet zich sterk af tegen Chalmers’ zombie-argument. Elk kenbaar aspect van een zombie is identiek aan die van een mens, en dus is het logischerwijs onmogelijk om een zombie te herkennen als wij deze zouden tegenkomen. Chalmers’ argument, dat het logisch denkbaar is dat een dergelijke zombie bestaat, is dus even aannemelijk als Dennetts argument dat wij een dergelijke zombie als drager van een bewustzijn zouden beschouwen. Dennett ziet Chalmers’ theorie derhalve als logisch incoherent. Er is namelijk geen manier om te kunnen bevestigen dat deze zombie geen bewustzijn heeft, dit kunnen we slechts aannemen. En doordat wij een zombie nooit van een echt persoon kunnen onderscheiden, is Chalmers’ argument irrelevant en zinloos. Dit geeft Dennett de ultieme sterkere positie, en zijn verdediging van het fysicalisme blijft vooralsnog gegrond.

Dennett en Chalmers’ discussie heeft gevolgen voor AI. Wanneer het fysicalisme namelijk onwaar zou zijn, levert dit implicaties op. Doordat de artificiële neurale netwerken die bij AI worden toegepast, berusten op slechts fysiek waarneembare processen, zou een dergelijk netwerk niet kunnen

resulteren in bewustzijn. Een kunstmatig intelligente machine zou in dit geval altijd Chalmers' definitie van een zombie zijn. Is het fysicalisme wel waar, dan levert dit in ieder geval de kennis op dat een AI-machine, al is het slechts in theorie, bewustzijn kan bezitten. Gezien de huidige ontwikkelingen, zoals *deep learning*, is dit een uiterst vruchtbare positie. Alhoewel bij *deep learning* nog geen sprake is van bewustzijn, kan dit desalniettemin bijdragen aan een stap in de richting van subjectieve ervaring bij AI.

4.2 De Volgende Stap

Dennetts bewustzijnstheorie berust voor een groot deel op evolutionaire ontwikkeling. Zo veronderstelt hij in 1995 dat de natuur slechts een complex algoritme is, en dat de mens een samenstelling is van voorgeprogrammeerde – noem het gedetermineerde – deeltjes. Willen wij dit kunnen begrijpen, dan moeten we door middel van *reverse engineering* bij de kern van de evolutie komen. Hetzelfde geldt volgens Dennett bij een kunstmatig systeem, die veelal zo ontwikkelt dat dit niet meer te begrijpen is voor de mens. Volgens Dennett moeten we eerst doen, en dan pas denken en terug redeneren; slechts dan kunnen we de werking van een AI-machine begrijpen. Dennett (2017a) lijkt hier echter op terug te komen, en stelt zelfs dat het onmogelijk is om evolutie, en in het verlengde AI-algoritmen, te kunnen begrijpen. Het is belangrijk om op te merken dat Dennett evolutie stapsgewijs ziet in dit specifieke geval: van kern – de eerdergenoemde robot – naar mens en vervolgens naar AI. Het gaat Dennett in *From Bacteria to Bach and Back* niet om de sprong van kern naar mens, maar van mens naar AI. Doordat het bewustzijn hier een belangrijke rol speelt, en gezien onze geringe kennis hiervan op dit moment, is het voor Dennett niet mogelijk om te begrijpen hoe deze stap zich voltrekt. Dit doet denken aan de eerdergenoemde black box die ook bij het connectionisme resulteert in een beperkt begrip van de werking van AI.

Het Cog project, waar Dennett aan heeft deelgenomen, was een ultieme poging deze black box te ontmantelen. De bedoeling was dat Cog vanuit zijn meest primitieve functies zou uitgroeien tot een intelligente – en wellicht zelfs bewuste – robot, ofwel een humanoïde androïde. Voor Dennett zou een dergelijke ontwikkeling filosofisch van belang zijn om zo de mechanische processen, die voor verandering en ontwikkeling zorgen, te kunnen analyseren en begrijpen. Alhoewel dit niet is gelukt, betekent dit niet dat het project tevergeefs was. Zo zijn er door Dennett enkele voorwaarden opgesteld voor het hebben van een bewustzijn bij AI, die al dan niet van toepassing zijn voor menselijk bewustzijn: sensorische waarneming, belichaming en motoriek, en als meest belangrijk linguïstiek vermogen. In het huidige filosofisch en cognititie-wetenschappelijk debat over het bewustzijn zijn deze criteria terug te zien: bewustzijn wordt hedendaags getypeerd als kunnen ervaren en waarnemen, wat dan ook de enige consensus binnen het debat is (Van Gulick, 2018). Waarnemen en ervaren omvat echter Dennetts voorwaarden voor bewustzijn. Zo is sensorische waarneming immers toegepaste waarneming. Belichaming en motoriek beïnvloeden onze ervaring, en linguïstiek vermogen is een resultaat van gehoor (spreken) en zicht (schrijven).

Tot slot doet Dennett in *From Bacteria to Bach and Back* enkele uitspraken over de toekomstige omgang met AI-machines. Hij benadrukt hier dat AI nooit de rol van mens mag innemen; het moet bijvoorbeeld altijd duidelijk zijn wanneer een persoon met een kunstmatige machine spreekt. Deze transparantie dient tevens verlengd te worden naar de mogelijke risico's van AI. Dennett formuleert een ethiek die nooit eerder nodig is geweest, en beperkt zich mede hierdoor tot de basisnormen. Het is op dit moment dan nog niet mogelijk om de details van een ethiek over AI uit te werken, maar dat is des te meer reden om AI behoedzaam te benaderen. Hij is dan ook van mening dat het zo menselijk mogelijk maken van een intelligente robot een inefficiënte toepassing is van een potentieel efficiënt systeem. Of zoals Dennett zelf zegt, "we zijn veel beter af met ons gereedschap dan met onze collega's" (Dennett, 2017b). Het is dan ook hierom dat Dennett AI *an sich* niet beschouwt als 'gevaarlijk' maar veeleer als neutraal, de mens daarentegen is in staat haar te misbruiken. Er moeten dus risicobeperkende maatregelen genomen worden om een toekomst te kunnen garanderen waarin mens en AI-machine in symbiose kunnen bestaan.

4.3 De Synthese

Dennett was een van de eerste pioniers van een filosofisch onderlegde AI-theorie. Kunstmatig intelligente systemen zijn echter grotendeels afhankelijk van het bewustzijnsdebat; we kunnen pas echt een intelligent – en potentieel bewust – systeem creëren wanneer we een bevredigende definitie hebben van bewustzijn. De eerste grote stappen zijn inmiddels gemaakt, te beginnen bij het connectionisme. Dennetts voorspelling (1978), dat modellen op neuraal niveau zullen worden gemaakt, is dan ook verhelderend en terecht gebleken. Dit was voor hem belangrijk om te kunnen bepalen welke delen van het brein een kunstmatig intelligente machine dient te simuleren, en welke irrelevant zijn. Gezien het huidige optimisme omtrent *deep learning*, en de eerdergenoemde werking hiervan, is Dennetts analyse zelfs dertig jaar later nog aantoonbaar relevant. Ook is het frame-probleem van AI nog altijd onopgelost, wat aantoont dat Dennett wederom scherpe analyses heeft gemaakt, die AI-programmeurs nog altijd bezighouden.

Dennetts *multiple drafts* theorie tracht om voor eens en voor altijd van het idee van een Cartesiaans theater af te stappen. Zijn sterkste argument tegen een dergelijk dualisme is dat het bewustzijn te verklaren valt als een constante informatiestroom, die parallel het brein inkomt. Deze informatiestroom heeft geen aanwijsbaar begin- of eindpunt, en dit kan dan ook over het bewustzijn gesteld worden. Bewustzijn is bij Dennett niet meer dan een neurale verwerking en opslag van dergelijke informatiestromen; er is dus geen onverklaarde factor die het bewustzijn mogelijk maakt. Dit is in lijn met het fysicalisme, dat hedendaags voornamelijk terugkomt bij neuro- en cognitiewetenschap. Het moge duidelijk zijn dat Dennett hier niet impliceert dat het bewustzijn zelf materieel is, maar slechts dat het voortkomt uit materie. Dit wordt verduidelijkt aan de hand van het volgende voorbeeld: geluid is een immaterieel gegeven, het zijn slechts trillingen in de lucht. Het wordt echter veroorzaakt door

materie, namelijk twee of meerdere lichamen die met elkaar in contact komen. Dit is dus hoe het bewustzijn ook werkt volgens Dennett. Doordat meerdere entiteiten (cellen, neuronen) met elkaar in contact komen wordt er een nieuwe, immateriële entiteit gevormd. Op deze wijze is het argument van Dennett meer dan plausibel, en sluit het aan bij de hedendaagse, veelvoorkomende benadering van bewustzijn.

Tot slot houdt Dennett zich bezig met een evolutionaire benadering van het bewustzijn en AI. Alhoewel Dennett erkent dat AI zelf geen inherent evolutionair product is, komt zij toch voort uit evolutie. De mens is namelijk wel evolutionair tot stand gekomen, en is zodanig intelligent geworden dat een systeem gecreëerd kan worden dat nog vele malen intelligenter is. Er zit dus een bepaalde mate van transitiviteit in het evolutionaire proces dat tot AI leidt. Dennett is mede hierom van mening dat evolutie richtingloos en doelloos is; de onderliggende structuur die leidt van eencellig wezen tot AI-machine is zodanig complex dat dit onbegrijpelijk is. Dit lijkt echter tegengesproken te worden door de hedendaagse ontwikkeling van evolutionaire algoritmen, die het mogelijk maken om de evolutie te modelleren. Het is desalniettemin belangrijk om op te merken dat dit vooralsnog geen kritiek is op Dennett, maar slechts een toevoeging op zijn theorie. Want alhoewel deze algoritmen de evolutie in kaart brengen, begeven de onderliggende processen zich nog altijd in de veelbesproken black box. Wederom sluit Dennetts analyse bijna naadloos aan bij hedendaagse ontwikkelingen.

4.4 Conclusie

In dit hoofdstuk is duidelijk geworden dat Dennetts analyse van het bewustzijn en AI veelal accuraat en treffend is. Enkele verschillen die naar boven komen zijn, door middel van grondige inspectie, alsnog compatibel gebleken. Dennetts fysicalistische bewustzijnstheorie blijft gerechtvaardigd, zelfs na Chalmers' kritiek op basis van het zombie-argument. Dit is goed nieuws voor het vakgebied van AI: de aanname dat bewustzijn bij de mens voort kan vloeien uit louter materiële entiteiten, veronderstelt volgens Dennett dat dit bij AI ook het geval moet zijn. Zo kan in de toekomst een kunstmatig intelligent systeem een kunstmatig bewust systeem worden, al is het op dit moment maar in theorie. Dennett rekent dan ook af met iedere vorm van het lichaam-geest probleem. Het bewustzijn – en AI – stammen af van evolutie, en evolutie is niet meer dan een gecompliceerd algoritme, geproduceerd door de natuur. Zodoende is er sprake van een alomvattend, gesloten systeem volgens Dennett: door middel van evolutie kunnen wij AI begrijpen, en door middel van AI kunnen wij evolutie begrijpen. Desalniettemin moeten we niet op de zaken vooruitlopen, en dienen we behoedzaam om te gaan met AI. Want alhoewel AI op zichzelf genomen neutraal is, is de mens nog altijd in staat haar te misbruiken.

Conclusie

In deze thesis heb ik mij gericht op de vraag in hoeverre Dennetts analyse betreffende AI nog altijd relevant is, met betrekking tot hedendaagse ontwikkelingen. Het vakgebied van Artificiële Intelligentie is nog relatief nieuw, maar heeft tot op heden een turbulent bestaan geleid. Duidelijk is in ieder geval dat de opkomst van het connectionisme in de jaren 80 heeft gezorgd voor een aanpak door middel van artificiële neurale netwerken. Dit is tot op heden terug te zien bij *machine learning* en het latere *deep learning*. Dennett voorzag de toename van het gebruik van dergelijke netwerken eind jaren 70 al, en heeft vanaf dit moment aantoonbaar waardevolle bijdragen geleverd binnen het vakgebied van AI. Zo blijkt zijn fysicalistische en monistische bewustzijnstheorie bestand te zijn tegen kritiek, onder andere van David Chalmers. Dit levert een uiterst vruchtbaar toekomstperspectief binnen de discipline op, doordat het in theorie mogelijk is om een kunstmatige machine, naast intelligentie, ook over een bewustzijn te laten beschikken.

Het Cog project, waar Dennett aan heeft deelgenomen, is eveneens relevant. Alhoewel het project is gestrand, heeft het de definitie van bewustzijn bij een AI-robot verhelderd. Zo kan gesteld worden dat een dergelijke robot zintuigen moet hebben, niet slechts als software kan bestaan, en over een taalvermogen moet beschikken. Binnen het debat van cognitiewetenschap is dit waardevol, en dit kan worden verlengd naar de discipline van Kunstmatige Intelligentie. Dennett blijft echter kritisch, en stelt dat de mens te allen tijde behoedzaam om moet gaan met AI. Doen we dit niet, dan kan de mens haar misbruiken, met alle gevolgen van dien.

Een mogelijke beperking omtrent Dennetts theorie is dat het ook hem niet gelukt is om de black box van het bewustzijn, en in het verlengde daarvan bij AI, te openen. Hier is hij echter niet de enige in; tot nog toe is het geen enkele onderzoeker gelukt om op dit stuk een volledige transparantie te creëren. Dit heeft als gevolg dat zelfs *reverse engineering*, waar Dennett zelf voorstander van is, een uiterst gecompliceerde taak is. Volgens Dennett doen we er in ieder geval goed aan om te erkennen dat AI een product van de evolutie is, welke zelf algoritmisch functioneert. Dit kan ons helpen begrijpen hoe de structuur en hiërarchie van de natuur werkt.

Al met al moet er dus nog veel gebeuren, zowel binnen het filosofisch en cognitie-wetenschappelijk debat als binnen de discipline zelf. Een ding is echter zeker: Dennett heeft zich bewezen als een van de pioniers der AI, en zijn stem blijft tot op heden krachtig (en terecht) doorluiden.

Literatuurlijst

Aleksander, I. (1995). Artificial Neuroconsciousness: An Update. In J. Mira, & F. Sandoval (Reds.), *From natural to artificial neural computation : International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, June 7-9, 1995 : proceedings*. (pp. 566-583). New York, VS: Springer-Verlag.

Bennett, C. C., & Hauser, K. (2013). Artificial Intelligence Framework for Simulating Clinical Decision-Making: A Markov Decision Process Approach. *Artificial Intelligence in Medicine, 57*(1), 9-19. doi:10.1016/j.artmed.2012.12.003

Chalmers, D. J. (1995). *The Conscious Mind: In Search of a Theory of Conscious Experience*. Oxford, Groot-Brittannië: Oxford University Press.

Creath, R. (2017). Logical Empiricism. Geraadpleegd op 3 juni 2018, van https://plato.stanford.edu/archives/fall2017/entries/logical-empiricism/

Crevier, D. (1993). *AI: The Tumultuous Search for Artificial Intelligence*. New York, VS: BasicBooks.

Dennett, D. C. (1978). Artificial Intelligence as Philosophy and as Psychology. In D. C. Dennett (Red.), *Brainstorms: Philosophical Essays on Mind and Psychology* (pp. 109-126). New York, VS: Bradford Books.

Dennett, D. C. (1981). Where Am I? Reflections. In D. R. Hofstadter, & D. C. Dennett (Reds.), *The Mind's I: Fantasies and Reflections on Self and Soul* (pp. 230-231). New York, VS: Bantam Books.

Dennett, D. C. (1984). Cognitive wheels: the frame problem of AI. In C. Hookway (Red.), *Minds, Machines and Evolution* (pp. 129-150). Cambridge, Groot-Brittannië: Cambridge University Press.

Dennett, D. C. (1991). *Consciousness Explained*. New York, VS: Little, Brown and Company.

Dennett, D. C. (1995). *Darwin's Dangerous Idea*. New York, VS: Penguin Books.

Dennett, D. C. (1997). Consciousness in human and robot minds. In M. Ito, Y. Miyashita, & E. T. Rolls (Reds.*), Cognition, Computation and Consciousness* (pp. 17-29). New York, VS: Oxford University Press.

Dennett, D. C. (1997). Cog as a thought experiment. *Robotics and Autonomous Systems*, *20*(2-4), 251-256.

Dennett, D. C. (2017a). *From Bacteria to Bach and Back: The Evolution of Minds*. New York, VS: W. W. Norton & Company.

Dennett, D. C. [Big Think]. (2017b, 2 april). It Would Not Be Cool If AI Were Conscious — It Would Be Dumb | Daniel Dennett [YouTube Video]. Geraadpleegd op 3 juli 2018, van https://youtu.be/KkgxOXmF4zk

Dooremalen, H., De Regt, H., & Schouten, M. (2011). *Stof tot denken: Filosofische aspecten van brein en bewustzijn* (2e ed.). Amsterdam, Nederland: Boom.

Dreyfus, H. (1972). *What Computers Can't Do*. New York, VS: MIT Press.

Dreyfus, H. (1992). *What Computers Still Can't Do*. New York, VS: MIT Press.

Dreyfus, H., & Dreyfus, S. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Oxford, UK: Blackwell.

Fearn, N. (2007). *The Latest Answers to the Oldest Questions: A Philosophical Adventure with the World's Greatest Thinkers*. New York, VS: Grove Press.

Flusberg, S. J., & McClelland, J. L. (2014). *Connectionism and the Emergence of Mind.* doi:10.1093/oxfordhb/9780199842193.013.5

Garson, J. (2016). Connectionism. Geraadpleegd op 3 juni 2018, van https://plato.stanford.edu/archives/win2016/entries/connectionism/

Kato, I. (z.d.). Humanoid History -WABOT-. Geraadpleegd op 14 juli 2018, van http://www.humanoid.waseda.ac.jp/booklet/kato_2.html

Levin, J. (2017). Functionalism. Geraadpleegd op 4 juni 2018, van https://plato.stanford.edu/archives/win2017/entries/functionalism/

McCarthy, J. M., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955. *AI Magazine, 27*(4), 12-14.

McLelland, J. L. (2010). Emergence in Cognitive Science. *Topics in Cognitive Science, 2*(4), 751-770. doi:10.1111/j.1756-8765.2010.01116.x

Mooney, R. J. (1997). Inductive logic programming for natural language processing. *Inductive Logic Programming - 6th International Workshop, ILP-1996, Selected Papers*, *1314*, 3-22.

Rescorla, M. (2017). The Computational Theory of Mind. Geraadpleegd op 21 juni 2018, van https://plato.stanford.edu/archives/spr2017/entries/computational-mind/

Searle, J. (1980). Mind, brains, and programs. *The Behavioral and Brain Sciences, 3*, 417-457.

Shannon, C. E. (1950). Programming a Computer for playing Chess. *Philosophical Magazine, 41*(7), 256-275.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 49*, 433-460.

Van Gulick, R. (2018). Consciousness. Geraadpleegd op 2 juli 2018, van https://plato.stanford.edu/archives/spr2018/entries/consciousness/