# De creatie van zelfbewuste machines

De mogelijkheid van kunstmatig zelfbewustzijn
in historisch perspectief

Eburon
Utrecht 2021

ISBN 978-94-6301-370-3

Academische Uitgeverij Eburon, Utrecht
www.eburon.nl

Omslagontwerp: Textcetera, Den Haag

# Inhoud

# Voorwoord

Zal de mens een oude droom waarmaken en uiteindelijk in staat zijn zelfbewuste machines te creëren? In de geschiedenis van de mensheid is de ontwikkeling van kunstmatige intelligentie van niet te overschatten importantie. Zal die ontwikkeling uitlopen op de creatie van machines, die net als de mens zich van zichzelf bewust kunnen zijn? En wat zou een dergelijke uitzonderlijke ontwikkeling betekenen in het licht van de hele geschiedenis van mens en wereld? Het zijn deze fundamentele vragen waarop in dit boek via strenge analyses een antwoord wordt gezocht.

Het onderzoek, waarvan dit boek het resultaat is, is de voortzetting van werk dat jaren geleden is begonnen. In dat onderzoek speelt een relationele systeemtheorie een centrale rol. Die systeemtheorie is een metatheorie, want het is een theorie die algemeen de relaties tussen systemen en de consequenties daarvan analyseert en daarom op alle gebieden van de werkelijkheid, waar we met systemen te maken hebben, toegepast kan worden en zo systeem-specifieke theorieën kan verhelderen. Deze metatheorie, de theorie van de vier fundamentele systeemrelaties, maakt het mogelijk de grondstructuren die de disciplines op hun gebied aan het licht hebben gebracht inzichtelijk te maken.

Een aanzet tot deze systeemtheorie heb ik gebruikt in mijn wetenschapsfilosofisch proefschrift "*Science and Liberation*"[1] met het oog op grondslagenonderzoek van de moderne wetenschap in het algemeen en van de natuurkunde in het bijzonder.

Deze metatheorie heb ik verder ontwikkeld in mijn boek "*Verheldering van de werkelijkheid*"[2]. Hoofdthema van dat boek is de toepassing van deze systeemtheorie op een aantal heel verschillende wetenschapsgebieden. Dat boek bevat zo onder andere ook een hoofdstuk waarin onderzocht wordt in hoeverre deze metatheorie op het gebied van de biologie toegepast kan worden, namelijk om de structuurverschillen tussen plant, dier en mens te verklaren. Het zijn vooral de in dat onderdeel uitgevoerde analyses die nu worden voortgezet en ik gebruik in deze studie dan ook zo nu en dan materiaal aan dit boek ontleend.

In deze nieuwe publicatie is de centrale vraag of zelfbewuste machines gecreëerd kunnen worden. Daarvoor is nodig te weten wat het zelfbewustzijn van de mens inhoudt. Een primaire vraag is daarom wat de structuur is die aan zelfbewustzijn ten grondslag ligt. Met behulp van de metatheorie lijkt het mogelijk deze vraag te beantwoorden. Daarmee kan dan ook een antwoord verkregen worden op wat wel het 'moeilijke probleem' van bewustzijn wordt genoemd: "how mind can arise from matter".

Na het onderzoek of het creëren van een zelfbewuste machine in principe mogelijk is, zal ook moeten worden nagedacht over wat deze ontwikkeling tot gevolg zal of kan hebben en wat het in wijder verband voor de toekomst van mens en wereld betekent. We kunnen niet om deze

1. *Science and Liberation. A blind spot in scientific research – exploring a new structure of reality.* Thesis Publishers, Amsterdam, 1991.
2. *Verheldering van de werkelijkheid. Inzicht in de ontwikkeling van wetenschap en samenleving middels een relationele systeemtheorie.* Het Zuiden, Vught, 2011.

vraag heen, ook omdat in onze tijd – veroorzaakt door de mens zelf – de mens en zijn toekomst door dodelijke gevaren worden bedreigd: het bestaan van kernwapens en de opwarming, vervuiling en uitputting van de aarde. Deze studie eindigt daarom met een reflectie over wat de mogelijke ontwikkeling van zelfbewuste machines in wereldhistorisch perspectief zou kunnen betekenen.

# Inleiding

## 1. De oude en nieuwe droom van een kunstmatig mens

Kan een machine geconstrueerd worden die zich net als de mens bewust is van zichzelf? Is het mogelijk een robot te maken die over zelfbewustzijn beschikt en die over zichzelf en de wereld kan nadenken? En als dat mogelijk is, wat zal de realisering daarvan dan voor de mensheid betekenen? Deze vragen komen in onze tijd met volle kracht op, zeker nu sinds het midden van de vorige eeuw de ontwikkeling van kunstmatige intelligentie een enorme vlucht heeft genomen.[3]

De gedachte dat de mens kunstmatig uit materie een mensachtig wezen zou kunnen maken is niet nieuw. Al eeuwen geleden is dat idee bij mensen opgekomen en heeft men over de daarmee gepaard gaande dilemma's en gevaren nagedacht. Uit de tijd van het vroege jodendom (uit de Talmoed) stamt de legende van de golem, een mensachtig wezen dat net als Adam uit klei geschapen is. Tot deze joodse traditie horen ook de verhalen over de 16e-eeuwse rabbi Löw uit Praag die een golem gecreëerd zou hebben. Eeuwenoud is ook het alchemistisch project om uit dood materiaal een 'homunculus', een artificiële mens, te creëren. Die gedachte vinden we ook in Goethes Faust. In 1818 verscheen Mary

3. Onder "kunstmatige intelligentie" wordt zoals gebruikelijk verstaan "a computer program developed by humans that can achieve complex goals". Zie B.P.Göcke, A.Rosenthal-von der Pütten (Eds.) *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*, Brill, Leiden, 2020, p.IX.

Shelley's befaamde boek *Frankenstein, or the Modern Prometheus*.[4] De auteur was een kind van de verlichting, maar de ambitie van haar hoofdpersoon Victor Frankenstein om een homunculus te maken, is in feite een droom uit een premoderne alchemistische tijd. De ondertitel van het boek duidt erop dat we te maken hebben met een soort herhaling van de Prometheus-mythe.

Moderne wetenschappelijke en technische ontwikkelingen hebben wezenlijk nieuwe voeding gegeven aan de gedachte dat het mogelijk moet zijn kunstmatig een mensachtig wezen te creëren. In 1950 publiceerde Turing zijn artikel "Computing Machinery and Intelligence".[5] Daarin voorzag hij de mogelijkheid van zelfdenkende intelligente machines. Sindsdien wordt er ten gevolge van de stormachtige ontwikkeling van kunstmatige intelligentie en het grote potentieel daarvan intensief nagedacht over de vraag wat deze technologische revolutie ons gaat brengen. In non-fictie maar vooral ook in (vaak sciencefiction achtige) romans en films figureren veelvuldig zelfbewuste artefacten en in dat verband worden allerlei toekomstbeelden overwogen. De wereld die dan geschetst wordt krijgt niet zelden het karakter van een dystopie.

Nu zijn er ook auteurs die de mogelijkheid van zelfbewuste machines, als een essentiële stap verder dan 'denkende machines', zonder meer ontkennen. Volgens hen zullen robots, hoe intelligent en zelflerend ze ook mogen worden, niet tot wezenlijk meer in staat zijn dan wat er door de makers aan algoritmen is ingestopt. Anderen daarentegen, met name aanhangers van wat 'strong Artificial Intelligence'(AI) wordt genoemd,

4. De grote invloed van dit boek blijkt wel uit de volgende cijfers: "There have been more than 300 editions of the original novel; more than 650 comic books and cartoon strips inspired by it; over 150 fictional spin offs and parodies; at least 90 films (..) and something like 80 stage adaptations." Richard Holmes, "Out of Control", in: *The New York Review of Books,* 21 december 2017, p.38.
5. Alan Turing, in: *Mind* (1950) 59, p.433-460.

beklemtonen dat er geen wet of principe is dat het bestaan van zelfbewustzijn en subjectieve gevoelens in door mensen ontworpen artefacten verbiedt. Een overweging daarbij is, dat in de loop van de evolutie de mens zelf is ontstaan als een hoeveelheid materie met een bijzondere structuur. Waarom zouden er dan geen systemen gecreëerd kunnen worden – nu niet gemaakt van vochtig organisch materiaal maar van silicium en staal – met een vergelijkbare, zelfbewustzijn mogelijk makende, structuur? De cruciale vraag is dan natuurlijk – een vraag die in deze studie centraal zal staan – welke structuur daarvoor noodzakelijk is.

Een andere kwestie is of een dergelijke ontwikkeling wel gewenst is. Zijn pogingen daartoe niet een vorm van hybris? En loopt de mensheid door hoogontwikkelde kunstmatige intelligentie niet een groot gevaar, bijvoorbeeld door de ontwikkeling van autonome wapensystemen of doordat artefacten zich tegen hun makers kunnen keren? Enkele jaren geleden schreef een groep experts een open brief waarin ze opriepen tot onderzoek naar de maatschappelijke impact van kunstmatige intelligentie. Ze erkenden natuurlijk de grote voordelen ervan, maar ze wezen er ook op dat een verdere ontwikkeling "out of control" zou kunnen raken. Bovenmenselijke kunstmatige intelligentie zou onberekenbare effecten kunnen hebben en zelfs het einde van de mensheid kunnen betekenen. Ze riepen op om bij de verdere ontwikkeling in ieder geval niet iets te creëren dat niet gecontroleerd kan worden.

Deze oproep lijkt in te houden dat deze experts in ieder geval de ontwikkeling van zelfbewuste machines als ongewenst beschouwen en deze bij voorbaat willen uitsluiten. Want zelfbewustzijn en de mogelijkheid om over zichzelf en de wereld na te kunnen denken betekenen – net als bij de mens – wezenlijk *vrijheid* en dus oncontroleerbaarheid. Maar zou dat echt een bedreiging zijn of kan het ook een positieve ontwikkeling zijn, een mogelijkheid tot nieuwe vrijheid?

Als een kunstmatige zelfbewuste machine een reële mogelijkheid is, dan is te verwachten dat die ook eens gerealiseerd zal worden. Dat zal

dan niet alleen belangrijk zijn voor inzicht in wie en wat de mens zelf is, maar ook in wijder verband enorme consequenties hebben. De vraag is met name wat dat zou betekenen in het licht van de huidige crises en gezien de uitdagingen waar de mensheid op dit moment voor staat. Zal het naast de gevaren die de mensheid nu al bedreigen een extra bedreiging gaan vormen of biedt het veeleer nieuwe mogelijkheden?

Zullen, zoals sommige transhumanisten denken, intelligente robots de aarde van de mensen erven en beter dan de mens in staat zijn het universum verder te verkennen? Moet zo'n ontwikkeling als een nieuwe stap in de evolutie beschouwd worden? Deze vragen wijzen er op dat de ontwikkeling van AI ook van belang is vanuit een nog algemener perspectief. Het ontstaan en de ontwikkeling van de moderne wetenschap en techniek is in de geschiedenis van de mensheid van buitengewone en niet te overschatten importantie. Dat de mens erin slaagt door rationeel methodisch onderzoek de grondstructuren van de werkelijkheid aan het licht te brengen en bovendien de zo verkregen kennis technisch te benutten is van wereldhistorische betekenis. In dat kader betekent het werk aan kunstmatige intelligentie zonder twijfel een uitzonderlijke nieuwe ontwikkeling. We kunnen er daarom niet omheen ook na te denken over wat het mogelijk ontstaan van zelfbewuste machines betekent in het licht van de gehele geschiedenis zowel van de mensheid als van het zich evoluerend heelal.

Al deze vragen en overwegingen stellen de moderne mens voor de taak om in een streng rationeel onderzoek na te gaan allereerst of een kunstmatig zelfbewuste machine in principe mogelijk is en vervolgens hoe deze dan gerealiseerd zou kunnen worden. Tenslotte moet dan ook overdacht worden wat de wijdere betekenis daarvan is.

## 2. Onderzoeksprogramma en methode

Vanwege de complexiteit en veelomvattendheid van deze problematiek is de uitvoering van dit onderzoeksprogramma geen eenvoudige opgave. Daarom moeten we eerst nagaan hoe dit te realiseren is.

Allereerst moet de vraag aan de orde komen of het realiseren van een zelfbewuste machine mogelijk is. Om te weten welke structuur een robot moet hebben om van een zelfbewuste robot te kunnen spreken, moeten we natuurlijk weten wat zelfbewustzijn is en vooral waardoor dat mogelijk wordt gemaakt. De enige manier om daar achter te komen is inzicht proberen te krijgen in het zelfbewustzijn zoals we dat bij de mens tegenkomen. Vandaar dat de primaire vraag is wat het zelfbewustzijn van de mens inhoudt, wat de fundamentele structuur is die eraan ten grondslag ligt en wat het mogelijk maakt.

Maar met deze vraag lijkt dit project direct al vast te lopen. Want het antwoord op deze vraag ontbreekt niet alleen maar lijkt ook nog heel ver weg. Er is intensief over het zelfbewustzijn in relatie tot de materiele basis nagedacht maar tot een bevredigend resultaat heeft dat niet geleid. Dat blijkt alleen al uit de volgende citaten uit het werk van auteurs die zich jarenlang met deze problematiek hebben bezig gehouden.

Dennett gaf zijn boek uit 1991 de titel “Consciousness explained”. Maar aan het eind van dit boek merkt hij op:

> *“My explanation of consciousness is far from complete. One might even say that it was just a beginning ..”.*[6]

6. Daniel C. Dennett, *Consciousness explained,* Penguin books 1991, p.455. Dennett overweegt ook de mogelijkheid van een zelfbewuste robot; zie Ch.14 §1: “Imagining a conscious robot.”

Searle schreef in 1997:

> "*The brain is a conscious machine. (..) But how about an artificial machine (..) could that be conscious? (..) Because we do not know how real brains do it, we are in a poor position to fabricate an artificial brain that could cause consciousness.*"[7]

Een aantal jaren later, in 2011, publiceerde Searle een review-artikel met als titel "The mystery of consciousness continues". Hij begint als volgt:

> "*How do neurobiological processes in the brain cause consciousness? I think this is the most important question in the biological sciences today.*"

Hij moet echter constateren:

> "*One depressing feature of this entire research project is that it does not seem to be making much progress.*"[8]

Het is duidelijk dat we te maken hebben met een tot nu toe onopgelost probleem.[9] Dit wordt dan ook wel "het moeilijke probleem" van de lichaam-geest problematiek genoemd. Hoe is in te zien dat een bijzondere materiestructuur zelfbewustzijn en subjectieve ervaringen kan voortbrengen? Kunnen ooit neurale mechanismen worden gevonden

7. John R. Searle, *The mystery of consciousness,* New York, 1997, p.202.
8. John Searle, in: *The New York Review of Books,* June 9, 2011, p.50.
9. Zo ook Yuval N. Harari in 2016: "Nobody has any idea how a congeries of biochemical reactions and electrical currents in the brain creates the subjective experience of pain, anger and love." *Homo Deus*, London 2016, p.108.