Rebecca Hendriks
3029700
Leon de Bruin
12.562 woorden
21 augustus 2017

Kunnen machines denken?

Intentionaliteit, kunstmatige intelligentie en de Chinese kamer

Scriptie ter verkrijging van de graad “Master of arts” in de filosofie
Radboud Universiteit Nijmegen

Hierbij verklaar en verzeker ik, Rebecca Hendriks, dat deze scriptie zelfstandig door mij is opgesteld, dat geen andere bronnen en hulpmiddelen zijn gebruikt dan die door mij zijn vermeld en dat de passages in het werk waarvan de woordelijke inhoud of betekenis uit andere werken – ook elektronische media – is genomen door bronvermelding als ontlening kenbaar gemaakt worden.

Plaats: Nijmegen

Datum:19-07-2017

Inhoudsopgave

Samenvatting

De vraag hoe slim computers zijn houdt computerwetenschappers en filosofen al decennialang bezig. Een van de belangrijkste en meest besproken argumenten tegen de mogelijkheid van kunstmatige intelligentie is het gedachte-experiment van de Chinese kamer van de Amerikaanse filosoof John Searle. Hij betoogt dat het hebben van het juiste computerprogramma niet hetzelfde is als denken of intelligentie, omdat wat de menselijke hersenen doen veel meer is dan alleen informatieverwerking. In deze scriptie onderzoek ik of het argument van Searle hout snijdt, waarbij ik met name zal kijken naar zijn definitie van intelligentie in termen van intentionaliteit.

Inleiding

In het dagelijks leven worden we omringd door computers die het leven makkelijker voor ons maken. Ze denken voor ons na, rekenen voor ons en onthouden onze afspraken. De temperatuur in ons huis wordt geregeld door slimme software, we kijken televisie op een smart tv en communiceren via onze smartphones. Maar hoe slim zijn deze slimme apparaten? Is de manier waarop we praten over computers een metafoor of mogelijk werkelijkheid? Voeren ze domweg uit wat door mensen is geprogrammeerd, of zijn er computers die echt nadenken, nu of mogelijk in de toekomst?

Deze vragen zijn prangender dan ooit, nu computers steeds meer dingen kunnen waarvan we voorheen dachten dat alleen mensen die konden. Computers kunnen mensen verslaan met schaken en go, spellen waarbij niet alleen computatie maar vooral intuïtie een belangrijke rol speelt. Bovendien kunnen computers componeren, essays schrijven en beeldende kunst maken.[1] Deze ontwikkelingen in kunstmatige intelligentie roepen de vraag op hoe menselijk onze intuïties, creativiteit en intelligentie zijn.

[1] Laurens Verhagen, “Slimme computers: kunnen ze straks ook kunst maken?,” *Volkskrant*, 15-07-2017.

De vraag hoe slim computers zijn houdt wetenschappers en filosofen al decennialang bezig. Een van de belangrijkste en meest besproken argumenten tegen de mogelijkheid van kunstmatige intelligentie is het argument van de Chinese kamer van de Amerikaanse filosoof John Searle. Hij betoogt dat informatieverwerking en hetzelfde gedrag vertonen als een mens niet hetzelfde is als kunnen denken of intelligentie. Wat de menselijke hersenen doen is meer dan wat een computer doet. Searle legt hierbij intelligentie uit in termen van intentionaliteit. Voor hem is intentionaliteit een voorwaarde voor intelligentie.

Voor de tegenstanders van Searle in dit debat is het helemaal niet zo vanzelfsprekend dat machines niet kunnen denken. Aan de hand van Searles gedachte-experiment van de Chinese kamer en de kritiek hierop zal ik de argumenten voor en tegen de mogelijkheid van kunstmatige intelligentie verkennen en een antwoord zoeken op de vraag of machines kunnen denken.

1. De Chinese kamer

In dit hoofdstuk zal ik het argument van de Chinese kamer uiteenzetten. Ik zal uitleggen wat het doel van John Searle is met dit gedachte-experiment en welke standpunten hij wil aanvallen. In zijn artikel "Minds, Brains and Programs"[2] richt hij zich tegen wat hij het project van sterke kunstmatige intelligentie (KI) noemt. Sterke KI beweert twee dingen, namelijk dat een op de juiste manier geprogrammeerde computer cognitieve toestanden heeft, en dat deze door het simuleren van cognitie de menselijke cognitie kan verklaren.

Searle verzet zich met name tegen de eerste bewering, dat als het juiste programma op een computer draait, er van de computer gezegd kan worden dat het cognitieve toestanden en intentionaliteit bezit. Intentionaliteit is de eigenschap van mentale toestanden waardoor ze ergens in de buitenwereld op gericht zijn of over gaan. Dat kan een object zijn of een stand van zaken. Wat Searle het meest tegenstaat, is het idee dat wat de geest betreft, de hersenen er niet toe doen, omdat het programma van de geest op elke hardware (die krachtig en stabiel genoeg is) gerealiseerd kan worden. Volgens Searle is dit idee "ontstaan uit het samensmelten van functionalisme en kunstmatige intelligentie."[3]

De geest is volgens de aanhangers van sterke KI meervoudig realiseerbaar, dat wil zeggen dat de geest in de hersenen gerealiseerd kan worden, maar ook op een computer of in theorie zelfs op een systeem van waterpijpen. Het hele project van sterke KI valt of staat met de meervoudige realiseerbaarheid van de geest. Als de geest namelijk alleen in de hersenen gerealiseerd kan worden, dan is de eerste bewering van sterke KI en daarmee het hele project onmogelijk.

Het gedachte-experiment van de Chinese kamer kan ook worden gezien als een argument tegen het functionalisme. Het functionalisme stelt dat mentale toestanden uitsluitend worden bepaald door hun functionele rol, namelijk de causale relaties die ze hebben met andere mentale toestanden, zintuiglijke input en gedragsoutput.

---

[2] John R. Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences* 3, no. 03 (1980): 417–24. doi:10.1017/S0140525X00005756.
[3] John R. Searle, *The Rediscovery of the Mind* (Cambridge, Mass.: The MIT Press, 1992), 44.

Om aan te tonen dat de twee beweringen van sterke KI niet kloppen, verwijst Searle naar een programma bedacht door de Amerikaanse KI-theoreticus Roger Schank. Deze laatste heeft in 1977 een programma bedacht met als doel het menselijk vermogen om verhalen te begrijpen te simuleren. Dit vermogen houdt onder andere in dat mensen vragen kunnen beantwoorden over verhalen, zonder dat deze informatie expliciet in het verhaal wordt genoemd.

Bij een verhaal over een restaurant zoals: “Een vrouw ging naar een restaurant en bestelde een hamburger. Toen de hamburger opgediend werd, was de vrouw erg tevreden. Ze gaf voor haar vertrek de ober een aanzienlijke fooi.” Als een lezer de vraag wordt gesteld of de vrouw de hamburger heeft opgegeten, zou deze ‘ja’ antwoorden. Een mens of machine die dit verhaal begrijpt, zou deze vraag correct kunnen beantwoorden, ook al wordt het antwoord niet expliciet in de tekst gegeven.

Schank wil met dit programma twee dingen aantonen. We kunnen ten eerste letterlijk van de machine zeggen dat het de verhalen begrijpt en ten tweede dat de machine en het programma het menselijke vermogen om verhalen te begrijpen verklaart. Maar volgens Searle volgen deze conclusies helemaal niet uit het programma van Schank. Dit is wat hij wil laten zien door middel van het gedachte-experiment van de Chinese kamer.

Searle richt zich in zijn argumentatie vooral tegen dit programma Schank, maar zijn kritiek is breder toe te passen op bijvoorbeeld de turingtest. De turingtest is door de Britse wiskundige Alan Turing in 1950 bedacht als alternatief op de vraag of machines kunnen denken. Deze test is een spel waarbij een speler een gesprek voert met een gesprekspartner. De speler weet niet of deze gesprekspartner een mens of een machine is. Als een machine wordt aangezien voor menselijk, dan slaagt deze voor de turingtest.[4]

[4] Alan M. Turing, “Computing Machinery and Intelligence,” *Mind* 59 (1950): 433-60.

1.1 Het gedachte-experiment

Om te laten zien dat de twee hierboven genoemde conclusies niet volgen, stelt Searle een gedachte-experiment voor. Dit gedachte-experiment draait om een persoon in een kamer met een stapel Chinese teksten. De persoon in de kamer kent geen Chinees; de teksten hebben voor hem geen enkele betekenis en hadden wat dat betreft net zo goed onderdeel van een niet-bestaande taal kunnen zijn. Naast de eerste stapel Chinese teksten krijgt de persoon in de kamer een tweede stapel teksten samen met een set van regels in de taal die de persoon wel begrijpt (in het geval van Searle is dit Engels). In de regels staat hoe Searle de teksten van de ene stapel moet toepassen op de teksten van de andere stapel. Een andere manier om dit te zeggen is dat Searle de ene set van formele symbolen op een andere set formele symbolen kan toepassen. Verder krijgt Searle een derde stapel Chinese teksten opnieuw samen met instructies in het Engels. Nu kan hij de teksten van de derde stapel toepassen op de teksten uit de eerste en tweede.

De volgende stap in het gedachte-experiment is dat Searle bepaalde Chinese symbolen terug dient te geven volgens de regels van de derde stapel. Buiten de kamer staan mensen die hem de tekens geven en in ontvangst nemen. Deze mensen noemen de eerste stapel een ‘script’, de tweede stapel een ‘verhaal’ en de derde stapel ‘vragen’. De symbolen die Searle teruggeeft noemen ze de ‘antwoorden op de vragen’, de set Engelstalige regels is het ‘programma’. Het verhaal in kwestie gaat over een restaurant. Searle is hiervan niet op de hoogte, de mensen buiten de kamer wel. Naast deze taak moet Searle Engelse verhalen lezen en daar vragen over beantwoorden in het Engels. Deze verhalen en vragen ontvangt en geeft hij op dezelfde manier als de Chinese input en output. Na verloop van tijd is Searle zo goed geworden in het manipuleren van de Chinese symbolen dat de antwoorden die hij in het Chinees geeft, even goed zijn als de antwoorden die hij over de Engelse verhalen geeft. De programmeurs zijn bovendien ook dusdanig goed in het schrijven van het programma dat het buiten de kamer gezien niet evident is dat Searle het Chinees niet beheerst.

Het resultaat van het manipuleren van de Chinese symbolen en het begrijpen van de Engelse teksten is dus hetzelfde voor de buitenstaander. Het grote verschil is volgens Searle echter dat hij de Engelse verhalen wel begrijpt, maar dat de Chinese tekens geen betekenis voor hem hebben. In het laatste geval handelt Searle, zo zegt hij, hetzelfde als een programma dat draait op een computer. De beweringen van sterke KI waren juist dat een computer op dezelfde manier geprogrammeerd als Searle de verhalen begrijpt op de manier dat een mens dat zou begrijpen. Op basis van het gedachte-experiment kunnen deze beweringen volgens Searle niet gemaakt worden.

In het licht van dit gedachte-experiment kijkt Searle opnieuw naar de twee beweringen van sterke KI. Volgens Searle spreekt het voor zich dat hij niets snapt van de Chinese verhalen. Er is input en output die niet te onderscheiden is van een Chinese moedertaalspreker, maar het begrijpen ontbreekt. Als Searle in de Chinese kamer niets begrijpt van de verhaaltjes, begrijpt een computer met Schanks programma ook niets van de verhaaltjes, in welke taal dan ook. De eerste bewering, dat een computer letterlijk de verhalen begrijpt, gaat dus niet op.

De tweede bewering is dat het programma menselijk vermogen tot begrijpen verklaart. Deze bewering is ook niet houdbaar zegt Searle. De computer en het programma bieden geen voldoende voorwaarden voor het begrijpen. Ze functioneren wel, maar van begrijpen is geen sprake. Het punt van Searle is dat welke formele principes je ook in een computer stopt, het is nooit genoeg om ervoor te zorgen dat de computer iets begrijpt, aangezien een mens dezelfde formele principes kan volgen zonder er iets van te begrijpen. Zulke principes zijn tegelijkertijd ook niet nodig, aangezien een persoon Engels kan begrijpen zonder een programma.

Drie illusies

Searle wijst op drie illusies die zorgen dat er veel mensen geloven dat sterke KI mentale fenomenen lijkt te reproduceren en daarmee uitlegt.[5] De eerste illusie is

[5] Searle, “Minds, Brains and Programs,” 423.

de verwarring die bestaat over het begrip informatieverwerking. Veel mensen denken dat mensen met hun geest informatie verwerken. Het lijkt dus zo te zijn dat als op een machine hetzelfde programma wordt gedraaid als in de hersenen, dat de informatieverwerking in beide gevallen identiek is.

Het probleem is dat wanneer mensen informatie verwerken, ze iets anders doen dan computers. Wat Searle wil aantonen met zijn gedachte-experiment is precies dat mensen informatie begrijpen, maar een computer niet. Wat een computer is doet is formele symboolmanipulatie. Volgens Searle kunnen we twee dingen doen. Of we definiëren informatieverwerking zodanig dat het intentionaliteit als onderdeel van het proces impliceert, of dat het dit niet doet. Als we voor de tweede optie kiezen, doet een computer niet aan informatieverwerking, maar alleen aan formele symboolmanipulatie op dezelfde manier dat een rekenmachine en thermostaat dit doen. Het is in dit geval aan een waarnemer van buitenaf om de input en output als informatie te interpreteren. In beide gevallen doet een computer niet hetzelfde als de hersenen wat betreft het verwerken van informatie.

De tweede illusie die Searle uit de weg wil helpen is wat hij noemt de "restjes van behaviorisme".[6] Omdat computers als ze goed zijn geprogrammeerd input-outputpatronen hebben die lijken op die van de mens. Een computer vertoont dan hetzelfde gedrag als een mens zou vertonen in reactie op stimuli van buitenaf. Hierdoor zijn we geneigd om intentionaliteit en mentale toestanden aan computers toe te schrijven. Zodra we weten dat een computer menselijke eigenschappen kan hebben zonder enige intentionaliteit, schrijven we er geen mentale toestanden meer aan toe.

De turingtest is een goed voorbeeld van een test die behavioristisch is: een computer slaagt voor de test als hij zich gedraagt als een mens. Het simuleren van menselijk gedrag is genoeg om als denkend gezien te worden. De verwarring die nu bestaat tussen simulatie en duplicatie zou volgens Searle niet meer bestaan als KI-wetenschappers behaviorisme af zouden wijzen.

---

[6] Ibid.

De derde illusie hangt samen met het lichaam-geestdualisme. Sterke KI maakt de aanname dat wat de geest betreft, de hersenen er niet toe doen. Wat er toe doet is het programma, het programma is onafhankelijk van hun realisatie in een machine. De geest is dus onafhankelijk van haar realisatie in de hersenen. Searle gelooft dat mentale fenomenen kenmerken zijn van de hersenen. Het project van sterke KI kan alleen bestaan als de geest los te koppelen is van de hersenen. Slechts op basis van deze aanname bestaat er de hoop om het mentale te reproduceren door middel van het schrijven en draaien van programma's. Deze vorm van dualisme beweert dus dat wat mentaal is aan de geest geen intrinsieke verbinding heeft met de hersenen.

Intrinsieke en afgeleide intentionaliteit

Om goed te begrijpen waarom Searle denkt dat computers geen intentionaliteit bevatten, moeten we eerst begrijpen wat Searle verstaat onder intentionaliteit en wat het verschil is tussen intrinsieke en afgeleide intentionaliteit. In *Intentionality: An Essay in the Philosophy of Mind*[7] legt hij dit uit. Intentionaliteit is zoals gezegd de eigenschap van mentale toestanden waardoor ze gericht zijn op iets in de buitenwereld en ergens over gaan.

Om uit te leggen wat de relatie is tussen een intentionele toestand en datgene waar die toestand op gericht is, legt Searle intentionaliteit uit als representatie. Intentionele toestanden representeren een object of stand van zaken, op dezelfde manier dat taalhandelingen dat doen. De uitspraak "het regent" representeert een stand van zaken in de wereld, namelijk dat het regent. De intentionele toestand "ik geloof dat het regent" representeert op dezelfde manier een stand van zaken. Hiermee bedoelt Searle niet dat intentionele toestanden essentieel of noodzakelijk linguïstisch zijn. Baby's hebben namelijk wel intentionele toestanden, ook al kunnen ze dit nog niet in taal uitdrukken.

Hoe komt het dat we in termen van begrijpen praten als het om computers en computerprogramma's gaat? Searle zegt dat dit komt omdat we onze eigen

[7] John R. Searle, "Intrinsic Intentionality," *Behavioral and Brain Sciences* 3, no. 03 (1980):1-36. doi:10.1017/S0140525X00006038.

intentionaliteit uitbreiden naar de hulpmiddelen om ons heen.[8] Een automatische deur 'begrijpt' dat hij open moet als er iemand voor de deur staat, een thermostaat 'weet' wanneer de verwarming aan of uit moet. Dit is echter alleen een wijze van spreken en is filosofisch niet vol te houden. De manier waarop een apparaat iets begrijpt is metaforisch en niet de letterlijke manier waarop een mens iets begrijpt wanneer hij bijvoorbeeld vragen over een verhaal beantwoord.

Deze uitbreiding van intentionaliteit naar de dingen en vooral machines om ons heen, maakt de noodzaak duidelijk voor een verscherping van het begrip intentionaliteit. In "Intrinsic intentionality",[9] de reactie van Searle op de commentaren op zijn artikel, verduidelijkt hij dit onderscheid, wat hij in "Minds, Brains and Programs" niet expliciet heeft gedaan. Alleen mentale toestanden bezitten intrinsieke intentionaliteit. Afgeleide intentionaliteit is een manier om over dingen, zoals thermostaten, te praten. Deze vorm van intentionaliteit is altijd afhankelijk van de waarnemer. Een computer is afhankelijk van een mens om mentale toestanden toegeschreven te krijgen.

Elke intentionele toestand bestaat uit een representatieve inhoud in een bepaalde psychologische modus. Geloven, willen en vrezen zijn voorbeelden van psychologische modi. Wat karakteristiek is voor toestanden die intrinsieke intentionaliteit hebben is dat de representatieve inhoud en de psychologische modus niet van elkaar gescheiden kunnen worden. Dit kan alleen bij afgeleide intentionaliteit. In dit laatste geval is er alleen representatieve inhoud, maar die heeft een 'gebruiker' nodig die door middel van een specifieke psychologische modus in een bepaalde relatie tot deze inhoud staat en deze kan representeren. Bij intrinsieke intentionaliteit kan er geen aparte gebruiker zijn omdat de representatieve inhoud niet los kan worden gezien van de intentionele toestand.

Dit onderscheid tussen de twee soorten intentionaliteit dient niet alleen ter verdediging van het argument van Searle, maar ook als aanval op het functionalisme. Searle zegt dat het functionalisme is gebaseerd op het onvermogen dit onderscheid in te zien. Een computer kan misschien functioneel

[8] Searle, "Minds, Brains and Programs," 419.
[9] Searle, "Intrinsic Intentionality," 450-456.

identiek zijn aan een mens, maar omdat een computer geen intrinsieke intentionaliteit bezit zal een computer geen mentale toestanden hebben en niet kunnen begrijpen.

Searles conclusie

Searle staat niet helemaal negatief tegenover het idee van machines die Engels of Chinees kunnen begrijpen. Hij ziet het echter niet gebeuren als machines alleen worden gedefinieerd als iets waar een programma op wordt gedraaid. Searle ziet de menselijke hersenen als een speciale soort machine. Hij zegt dat alleen een machine kan denken, maar alleen machines met dezelfde causale eigenschappen als onze hersenen.[10]

Intentionaliteit, en dus begrijpen, wordt niet veroorzaakt door het draaien van een computerprogramma. Een computersimulatie van het formele proces van lactatie produceert geen melk, waarom zouden we geloven dat een computersimulatie van de hersenen intentionaliteit produceert?[11] Intentionaliteit wordt volgens Searle veroorzaakt door onze biologische samenstelling. Daardoor zijn we in staat om waar te nemen, te handelen, te begrijpen, te leren, etc. Het is in principe mogelijk dat andere natuurkundige en scheikundige processen intentionaliteit kunnen voortbrengen, maar waar het om gaat is dat intentionaliteit altijd wordt veroorzaakt door biologische fenomenen. Searle wijst dus niet per definitie een biologische computer af, zolang deze maar de juiste causale eigenschappen heeft.

Bij de turingtest is het voldoende dat een computer alleen maar doet alsof hij een mens is. Als het de ondervragers als het ware succesvol voor de gek kan houden, slaagt de machine. Volgens Searle is het simuleren van de geest echter niet genoeg. Een computer moet intentionaliteit en mentale toestanden hebben, voordat we kunnen zeggen dat een computer denkt.

---

[10] Searle, “Minds, Brains and Programs,” 424.
[11] Ibid., 426.

1.2 Drie reacties

In zijn artikel “Minds, Brains and Programs” bespreekt Searle een zestal reacties op het gedachte-experiment van de Chinese kamer. Drie van deze reacties en de manier waarop Searle ze behandelt zal ik in deze sectie bespreken, omdat ze mijns inziens de scherpste reacties zijn. Dit zijn de 1) de systeemreactie, het idee dat het niet gaat om de persoon in de kamer, maar om het gehele systeem 2) de robotreactie, waarbij de persoon de controle heeft over een robot in plaats van een kamer en 3) de combinatiereactie, het idee dat de combinatie van de systeemreactie en de robotreactie samen het argument van de Chinese kamer ontkrachten.

Systeemreactie

De systeemreactie is het eerste antwoord op de Chinese kamer dat Searle behandelt. Hij omschrijft deze kritiek als volgt.[12] Het individu begrijpt de Chinese verhalen misschien niet, maar hij is slechts onderdeel van het systeem. Het systeem in het geheel begrijpt de verhalen wel. Het is voor het individu in de kamer niet nodig om Chinees te begrijpen, het systeem waar hij deel van uitmaakt begrijpt het wel.

Deze kritiek verwerpt Searle met een variatie op zijn gedachte-experiment. Stel dat we Searle het hele systeem laten internaliseren. Dit houdt in dat alle regels en databanken met Chinese symbolen niet langer op papier staan, maar in zijn hoofd zitten. Searle zit niet meer in de kamer, de kamer zit in zijn hoofd. Searle houdt nog steeds vol dat hij geen Chinees zou begrijpen, met als gevolg dat het systeem ook geen Chinees begrijpt. Het systeem kan geen Chinees begrijpen omdat het systeem onderdeel is van Searle, die geen Chinees begrijpt.

Wat de aanhangers van de systeemreactie beweren is volgens Searle het volgende. Er zijn twee subsystemen in de man aanwezig. Zoals Searle hierboven al zegt, begrijpt de man in de geïnternaliseerde versie van het gedachte-experiment geen Chinees. De man *als* formeel symboolmanipulatiesysteem zou

[12] Ibid., 419-20.

echter wel Chinees begrijpen. De twee subsystemen zijn eentje die Engels begrijpt en eentje die Chinees begrijpt. Het grote verschil tussen deze twee subsystemen zit volgens Searle in intentionaliteit. Het Engelse subsysteem snapt waar de verhalen en antwoorden over gaan. Het Chinese subsysteem weet alleen wat hij moet doen met de formele symbolen die hij krijgt voorgeschoteld, maar niet waar deze over gaan. Het Engelse subsysteem heeft intentionaliteit, het Chinese subsysteem niet.

Voor de buitenstaanders is er geen verschil tussen de subsystemen, beide geven ze de juiste antwoorden op de vragen. Waar het Searle echter om gaat is dat formele symboolmanipulatie niet genoeg is om te kunnen zeggen dat een systeem begrijpt. Het maakt voor het punt van Searle uiteindelijk niet uit of Searle alle regels internaliseert of in de Chinese kamer zit, hij begrijpt hoe dan ook geen Chinees.

Als het allemaal zo onwaarschijnlijk is wat de aanhangers van de systeemreactie beweren, waarom beweren ze het dan? Veel mensen die geloven in de "ideologie"[13] van sterke KI opperen namelijk een vorm van deze kritiek. De enige reden die Searle kan bedenken is dat Searle dezelfde input en output heeft als een Chineesspreker. Dit is niet genoeg voor het begrijpen van Chinees. Andere mensen overtuigen dat hij Chinees kan en de turingtest halen betekent niet dat Searle echt Chinees begrijpt.

Later in zijn tekst, in de verwerping van een andere kritiek op de Chinese kamer, komt Searle terug op het zogenaamde restje behaviorisme.[14] Aangezien computers die op de juiste manier zijn geprogrammeerd input en output hebben die veel op die van een mens lijkt, zijn we geneigd om dezelfde mentale toestanden aan computers toe te schrijven als die we aan mensen toeschrijven. Maar omdat het mogelijk is om een systeem te ontwerpen met menselijke capaciteiten, maar zonder intentionaliteit, zouden we deze neiging moeten kunnen overwinnen, aldus Searle. Bij het toeschrijven van intentionele toestanden aan

[13] Ibid., 419.
[14] Ibid., 423.

computers moeten we goed onthouden dat het om afgeleide intentionaliteit gaat, geen intrinsieke intentionaliteit.

Robotreactie

De robotreactie is de tweede reactie op het gedachte-experiment van de Chinese kamer die Searle behandelt.[15] Deze reactie behelst een aanpassing van het experiment zodat Searle niet langer in een kamer zit, maar een robot bestuurt. Deze robot kan waarnemen door middel van camera's en handelen door middel van robotledematen. Dit alles wordt geregeld via de 'hersenen' van de robot. De aanhangers van deze reactie beargumenteren dat zo'n robot wel Chinees begrijpt (en ook andere mentale toestanden heeft). Deze robot staat namelijk, in tegenstelling tot de oorspronkelijke Chinese Kamer of de aangepast versie uit de systeemreactie, wel in contact met de buitenwereld en kan meer dan alleen passief vragen over een verhaal beantwoorden.

Searle brengt hier tegenin dat zijn tegenstanders met deze reactie toegeven dat voor het begrijpen van een taal meer nodig is dan het manipuleren van formele symbolen. Namelijk een causale relatie met de buitenwereld. Bovendien voegt de toevoeging van waarneming en motorische capaciteiten aan de robot geen begrip of intentionaliteit toe. Dit wordt duidelijk als we kijken naar wat Searle in dit geval doet. Hij zal meer Chinese symbolen en meer regels in het Engels moeten verwerken. Hij moet namelijk de input van de camera's in tekstvorm verwerken en de aanwijzingen voor de beweging van de robot doorgeven in tekstvorm. Searle weet nog steeds niet wat hij aan het doen is, of wat de gevolgen zijn van zijn handelingen voor de handelingen van de robot. De robot noch Searle heeft intentionele toestanden.

[15] Ibid., 420.

Combinatiereactie

De samenvoeging van de systeemreactie met de robotreactie heet de combinatiereactie. Searle omschrijft dit argument als volgt.[16] Als we een robot hebben wiens gedrag niet te onderscheiden is van het gedrag van een mens, die werkt als een geheel en niet als een computer met input en output, dan is het terecht om intentionaliteit aan het systeem toe te schrijven.

Searle noemt het inderdaad plausibel om intentionaliteit toe te schrijven aan zo'n systeem. Totdat we meer weten van het systeem, namelijk dat het een robot is dat slechts een programma draait. Zodra we een andere verklaring hebben voor het gedrag van de robot, namelijk dat de robot een computerprogramma draait, schrijven we geen intentionaliteit meer toe. We weten dan dat de robot niet meer is dan een "ingenieuze mechanische pop".[17] De enige plek waar intentionaliteit te vinden is, is bij Searle, omdat hij een mens met intentionele toestanden is. Maar hij ziet niet wat de camera's van de robot zien, heeft niet de intentie de robotarm te bewegen en begrijpt niet wat er door of tegen de robot gezegd wordt. De intentionele toestanden van Searle hebben niets te maken met de handelingen of waarnemingen van de robot of waar de Chinese teksten over gaan en hebben zo niets te maken met enige intentionaliteit van de robot zelf.

Er zijn voor Searle twee redenen om intentionaliteit toe te schrijven aan een systeem dat geen mens is. Hij noemt als voorbeeld primaten en huisdieren. We vinden het normaal om over deze dieren te praten als intentioneel omdat we zo hun gedrag kunnen verklaren en voorspellen en omdat ze van ongeveer dezelfde materie als mensen zijn gemaakt. De combinatie van de coherentie van het gedrag van de dieren en de juiste causale materie die hieraan ten grondslag ligt, maakt dat we deze systemen als intentionele systemen beschouwen. Het gedrag van een computer is het gevolg van een formeel programma, dus heeft het niet de juiste causale eigenschappen veroorzaakt door de fysieke substantie die intentionaliteit veroorzaakt.

---

[16] Ibid., 421.
[17] Ibid.

2. Tegen de Chinese kamer

Het gedachte-experiment van de Chinese kamer is, zoals Searle het noemt, een aanval op de samensmelting van het functionalisme met kunstmatige intelligentie. In dit hoofdstuk zal ik de functionalistische argumenten tegen Searle uiteenzetten, met eventuele reacties van Searle hierop. Als eerste zal ik (nogmaals) kort uitleggen wat het functionalisme inhoudt en hoe dit belangrijk is voor sterke KI. Vervolgens zal ik vier filosofen aan het woord laten en hun argumenten tegen Searle uiteenzetten.

2.1 Functionalisme

Het functionalisme is de doctrine die mentale toestanden in termen van hun functie beschrijft. [18] Van een mentale toestand M kan worden gezegd dat hij de M-taak uitvoert. Pijn is bijvoorbeeld de interne toestand die de pijntaak, of pijnrol, uitvoert. De pijnrol wordt gedefinieerd in termen van input, output en interne verbindingen. De input zijn omstandigheden die de pijn veroorzaken, de output is het gedrag dat door de pijn wordt veroorzaakt en de interne verbindingen zijn de causale verbindingen tussen pijn en andere mentale toestanden. Mentale toestanden vervullen dus causale rollen.

Door mentale toestanden op deze manier te beschrijven kan meervoudige realiseerbaarheid worden verklaard. Het verklaren van meervoudige realiseerbaarheid biedt grote voordelen voor kunstmatige intelligentie. Mentale toestanden kunnen theoretisch gezien in mensen, maar ook in computers worden gerealiseerd. Laten we stellen dat er in mensen een bepaalde soort neurale activiteit is die de pijnrol vervult, zoals de stimulatie van c-zenuwvezels. Mensen hebben dan pijn als deze c-zenuwvezels gestimuleerd worden. Als bij een

[18] Ian Ravenscroft, "Functionalism," in *Philosophy of Mind: A Beginner's Guide* (New York: Oxford University Press Inc., 2005), 50-63.
Voor deze paragraaf heb ik gebruik gemaakt van *Philosophy of Mind* van Ian Ravenscroft.

marsmannetje een toestand van silicium heeft, of als een robot een anorganische toestand heeft die de pijnrol vervult, dan kunnen ook deze wezens pijn hebben.[19]

Een groot probleem voor het functionalisme is echter het verklaren van bewustzijn. Het argument van de Chinese kamer is een voorbeeld van een aanval op functionalisten die in sterke KI geloven. Een computer kan functioneel hetzelfde zijn als een mens, maar geen bewustzijn hebben.[20]

2.2 Georges Rey

In het artikel "What's Really Going on in Searle's "Chinese Room"" stelt de Amerikaanse cognitiefilosoof Georges Rey dat de kritiek die Searle zelf geeft en weerlegt niet de relevantie kritiek is op het gedachte-experiment van de Chinese kamer.[21] Wat John Searle namelijk wil bereiken in "Minds, Brains and Programs", zo zegt Rey, is het weerleggen van de beweringen van sterke KI. Het werkelijke probleem is volgens hem echter dat er überhaupt problemen zijn met betrekking tot het toeschrijven van mentale en semantische eigenschappen aan wat dan ook, niet alleen aan computers. Wat is de specificatie van precies wat het is dat begrijpt? Wat zijn de juiste causale verbindingen met de wereld? Dat zijn de vragen die Searle volgens Rey had moeten stellen.

Rey valt allereerst het standpunt van Searle aan dat een machine de juiste biologische causale eigenschappen moet hebben om te kunnen begrijpen. Wat is het precieze doelwit van Searle, zo vraagt Rey zich af? Sterke KI is namelijk veel sterker dan het werk van Schank waar Searle zich bij het gedachte-experiment op baseert. Searle zegt dat de sterke KI beweert dat de turingtest een voldoende voorwaarde voor begrijpen is. Terwijl dit slechts een inputoutput-equivalent is van

---

[19] Janet Levin, "Functionalism," in *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta: (Metaphysics Research Lab, Stanford University, 2016).
https://plato.stanford.edu/archives/win2016/entries/functionalism
Dit voorbeeld is ontleend aan de *Stanford Encyclopedia of Philosophy*.

[20] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996) 94-99. Het zombie-argument van David Chalmers toont aan dat het niet genoeg is voor een computer om functioneel qua gedrag hetzelfde te zijn als een mens, zo'n 'zombie' zou qualia missen.

[21] Georges Rey, "What's Really Going on in Searle's "Chinese Room"," *Behavioral and Brain Sciences* 3, no. 03 (1980): 169-85. doi:10.1007/BF00354586.

een mens. Sterke KI gaat echter verder omdat het een functionalistische theorie van de geest postuleert. Het gaat dus niet alleen om de juiste relatie tussen input en output, maar ook om de interne mentale toestanden van een systeem waar we mentale toestanden aan toeschrijven en de manier waarop deze aan input, output en andere mentale toestanden gerelateerd zijn.

De in- en output moet bemiddeld worden door de juiste interne toestanden, of, in het geval van sterke KI, door het juiste programma. Een systeem moet zich niet alleen hetzelfde gedragen als een mens, maar moet ook hetzelfde programma volgen. Deze voldoende voorwaarde is KI-functionalisme. Het gaat de functionalist niet om het juiste gedrag, maar om de vraag of wat er gebeurt binnenin de Chinese kamer functioneel-equivalent is aan wat er gebeurt binnen in een Chinees sprekend mens. Op de manier waarop Searle het beschrijft is het antwoord op deze vraag ‘nee’.

Wat Searle beschrijft met de Chinese kamer is een teletype-equivalent van het gedrag van een Chineesspreker, de output is alleen geschreven tekst. Maar wat de output is van de Chinese kamer, oftewel wat het gedrag is, is irrelevant voor sterke KI. Het gaat erom of wat er binnenin gebeurt functioneel equivalent is aan wat er in een Chineessprekend persoon gebeurt. Door weg te laten wat er vanbinnen de Chinese kamer gebeurt, vereenvoudigt Searle het functionalisme dat hij wil aanvallen tot het behaviorisme.

De enige regels die Searle toestaat zijn regels met betrekking tot de tekens. De vorm van het gedachte-experiment van de Chinese kamer kan geen non-verbale reacties geven of op non-verbale input reageren. Hierdoor zal Searle in de kamer voor geen enkele gedragstest slagen, laat staan dat het systeem functioneel equivalent is aan iemand die Chinees spreekt.

Om Searles voorbeeld een mogelijk tegenvoorbeeld te laten zijn tegen sterke KI past Rey het aan, vergelijkbaar met de robotreactie van Searle. Searle volgt in zijn eigen versie van de Chinese kamer alleen de regels die de Chinese tekens aan elkaar verbinden. Volgens Rey moet Searle ook de regels volgen van de andere programma’s die de KI-functionalist poneert. Het begrijpen van een taal houdt ook in dat men de talige tekens kan verbinden aan waarnemingen,

overtuigingen en verlangens. Dit zijn programma's voor onder andere problemen oplossen en besluiten maken. Deze programma's hebben allemaal te maken met de relatie tussen de mentale wereld en de buitenwereld.

Met deze aanpassingen van Rey wordt het volgens hem meteen minder duidelijk waarom Searle geen Chinees begrijpt, zelfs niet een beetje. Searle is nu wel degelijk in staat om iets te bepalen over wat de symbolen betekenen. Rey vraagt zich af welke andere causale eigenschappen er nodig zijn om Chinees te begrijpen, naast het volgen van semantische regels. De persoon in de kamer heeft de juiste biologie, het is namelijk een mens, en hij heeft het juiste programma. Searle blijft volhouden dat deze persoon geen Chinees begrijpt. Rey oppert dat het aan het biologische aspect ligt.

De regels die de persoon opvolgt bevinden zich namelijk buiten de persoon, hij moet ze in een boek opzoeken. De KI-functionalist maakt op dit punt volgens Rey geen beweringen over wat er in het hoofd van de persoon omgaat. Het begrijpen van Chinees gebeurt namelijk niet in deze persoon. Het is de kamer die een Chinees sprekend mens na probeert te doen, de persoon in de kamer is slechts de processor. Rey vergelijkt het toeschrijven van het begrijpen van Chinees aan de persoon in de kamer, als het toeschrijven van alle kenmerken van het Britse Keizerrijk aan koningin Victoria.[22] Het is mogelijk, maar het is op geen enkele manier noodzakelijk.

Om dit probleem van onderscheid tussen de kamer als geheel en de persoon in de kamer, de systeemreactie, te weerleggen, laat Searle de persoon in de kamer alle regels uit het hoofd leren. Is het echter nu nog steeds het geval dat de persoon dan alsnog geen Chinees begrijpt? Rey denkt dat dit niet zo is. Zolang Searle de regels bewust toepast en er uren over doet om een antwoord te geven, is er "slechts marginaal"[23] begrip, maar er is begrip. Wanneer de persoon het proces stroomlijnt zodat ze de betekenis van de tekens erdoorheen 'horen', begrijpen ze evengoed Chinees als elke andere Chinees spreker. Voor de KI-functionalist komt het uit het hoofd leren van alle regels erop neer dat de processor wordt

[22] Rey, "What's Really Going On," 174.
[23] Ibid., 175.

geprogrammeerd met het hele systeem. Hierdoor is de persoon functioneel niet te onderscheiden van elke andere Chineesspreker. De KI-functionalist zal nu zo ver gaan door te beweren dat het er niet toe doet dat de persoon een mens is, het uit het hoofd leren is genoeg.

Nu Rey Searles standpunt heeft weerlegd dat er voor begrip de juiste causale eigenschappen nodig zijn, wat zijn dan wel de noodzakelijke en voldoende voorwaarden voor begrijpen en andere cognitieve toestanden? Rey wil dit laten zien door het begrip ‘geloven’ als voorbeeld te nemen, omdat hij zegt dat wanneer je iets gelooft, je dit ook begrijpt. Hij stelt vervolgens dat een voldoende voorwaarde voor object x dat voldoet aan eigenschap F is dat x voldoet aan de wetten waaronder de term ‘F’ valt. Een voldoende voorwaarde voor de uitspraak dat mensen dieren zijn, is omdat mensen voldoen aan de wetten waaronder dieren vallen.

Vergelijkbaar is een voldoende voorwaarde dat x mentale toestanden heeft, dat x voldoet aan wetten waaronder het mentale valt. Rey noemt als voorbeeld van zo’n wet dat de Chinese kamer (c) hypotheses over zijn omgeving op een rationele wijze selecteert die betrouwbaar zijn. Vervolgens zal c hieruit logische gevolgtrekkingen maken en op basis hiervan voldoen aan de beschrijving van de handeling met de meeste voorkeur. Met andere woorden, c zal doen wat hij gelooft dat het beste resultaat zal bereiken.

Rey heeft het hier over c, dus het hele systeem, en niet de persoon in de Chinese kamer die slechts een onderdeel is. De KI-functionalist schrijft mentale toestanden toe aan het systeem als geheel, niet aan een bepaald subsysteem, laat staan de processor, zegt Rey. Voor de persoon in de kamer zijn de tekens nog steeds betekenisloos, maar dit komt niet omdat hij niet de juiste biologische eigenschappen heeft. De persoon in de kamer staat niet in de juiste geloofsrelaties tot die zinnen, of in de juiste causale relaties met de omgeving. Het systeem als geheel staat wel in deze relaties.

2.3 Douglas Hofstadter

In de *open peer commentary* "Reductionism and religion" haalt Hofstadter fel uit naar het artikel van Searle.[24] Hij noemt het in de eerste zin een "religieuze schimprede" en een van de woestmakendste artikelen die hij ooit in zijn leven heeft gelezen.[25] Het mag duidelijk zijn dat Hofstadter het niet eens is met Searle. Hofstadter is echter niet alleen negatief; hij uit ook begrip voor het probleem waar Searle mee worstelt, namelijk dat het moeilijk is om voor te stellen hoe het mentale voort kan komen uit hersencellen. In zijn stuk valt hij de intuïties van Searle aan en zegt hij dat Searle het gedachte-experiment verkeerd voorstelt om zo de lezer op het verkeerde been te zetten.

Om aan te tonen dat de intuïties van Searle geen reden zijn voor zijn conclusies dat het project van sterke KI onmogelijk is, haalt Hofstadter Searles voorbeeld van het systeem van met elkaar verbonden waterpijpen aan.[26] Elke sluis correspondeert met een synaps in de hersenen van iemand die Chinees begrijpt. Als deze sluizen goed bediend worden, volgt er een correct Chinees antwoord op elke Chinese vraag. Searle vraagt de lezer waar in dit systeem het begrip zit. Noch de man die de sluizen open en dicht zet, noch de waterpijpen zelf begrijpen Chinees. De combinatie van de man en het systeem van waterpijpen heeft evenmin begrip, zoals Searle al stelt in zijn reactie op de systeemreactie.

Searle vindt het absurd dat in een systeem van waterpijpen intentionaliteit kan bestaan. Hofstadter zegt dat alleen omdat het absurd is, dit niet betekent dat iets niet zo is of kan zijn. Het idee dat intentionaliteit, intelligentie, geest en bewustzijn voortkomen uit de hersenen klinkt absurd, maar toch is het zo. Als we dit eenmaal accepteren, is het ook niet raar om aan te nemen dat intentionaliteit uit een ander systeem kan ontstaan. Bovendien moeten we accepteren dat intentionaliteit het gevolg is van een formeel proces. De fysieke processen van de hersenen zijn volgens Hofstadter formeel, omdat ze regels volgen. Searle

[24] Douglas R. Hofstadter, "Reductionism and Religion," *Behavioral and Brain Sciences* 3, no. 03 (1980): 433-34. doi:10.1017/S0140525X00005847.
[25] Hofstadter, "Reductionism and Religion," 433.
[26] Searle, "Minds, Brains and Programs," 421.

accepteert dit echter niet, zegt Hofstadter. De kritiek van Hofstadter op de intuïties van Searle is uiteindelijk een kritiek op de overtuiging van Searle dat alleen de hersenen de juiste causale eigenschappen voor intentionaliteit bevatten.

Bovendien vinden we intentionaliteit niet op het niveau van deeltjes, maar op het niveau van het geheel. Zoals blijkt uit Searles reactie op de systeemreactie is hij het ook met deze overtuiging niet eens. De combinatie van deze twee stellingen, namelijk dat de intuïties van Searle en dat hij op het verkeerde niveau zoekt naar intentionaliteit, maakt het juist voor Hofstadter onvermijdelijk om toe te geven dat kunstmatige intelligentie mogelijk is.

Het tweede probleem dat Hofstadter heeft met het argument van de Chinese kamer is dat Searle ineens een enorme stap maakt in complexiteit. Eerst beantwoordt Searle alleen vragen over een restaurant, ineens slaagt hij voor de turingtest. Beide scenario's zijn implausibel. Hij verandert de tijdschaal waarin de handelingen plaatsvinden, een mens zou namelijk enorm lang doen over die processen. Bovendien is het niveau van omschrijving aangepast om zo de lezer erin te luizen dat het niet mogelijk is voor een programma om bewustzijn te hebben. Op deze manier geeft Searle een vertekend beeld van wat er werkelijk aan de hand is.

De vijf knoppen

In het boek *The Mind's I: Fantasies and reflections on self and soul* gaat Hofstadter opnieuw in op de Chinese kamer.[27] In deze publicatie hebben Daniel Dennett en Hofstadter essays en teksten over de geest en het zelf gebundeld samen met hun commentaren erop. Het commentaar dat Hofstadter in dit boek op Searle geeft is een versie van de systeemreactie.

Volgens Hofstadter moeten we Searles gedachte-experiment in een bredere context plaatsen om het beter te begrijpen. Dit soort experimenten wordt volgens hem namelijk gekenmerkt door de instellingen van vijf knoppen. Deze knoppen zijn:

[27] Douglas R. Hofstadter en Daniel C. Dennett, *The Mind's I: Fantasies and Reflections on Self and Soul* (Brighton: Harvester Press, 1981)

1. De fysieke materie waaruit de simulatie bestaat. In het geval van de Chinese kamer is dat een kamer met stukjes papier waarop symbolen en instructies staan.
2. Het niveau van nauwkeurigheid waarmee het de menselijke hersenen probeert te simuleren. In ons geval is dat het niveau van concepten en ideeën.
3. De fysieke grootte van de simulatie. De grootte is in dit geval een kamer.
4. De grootte en natuur van de demon die de simulatie uitvoert. De demon is Searle, een mens.
5. De snelheid waarop de demon werkt. De demon werkt langzaam, hij voert een handeling per paar seconden uit.

Om zijn punt te illustreren haalt Hofstadter er een tweede gedachte-experiment bij. Bij dit experiment van de Amerikaanse cognitiefilosoof John Haugeland zijn de vijf knoppen anders ingesteld. De instellingen van Haugeland zijn als volgt.

1. Neuronen en chemicaliën.
2. Niveau van het afvuren van neuronen.
3. De grootte van de hersenen
4. Een minuscule demon
5. Een duizelingwekkend snelle demon.

Wat Haugeland ons wil laten inbeelden is het werkelijke brein van een vrouw dat helaas defect is. Het kan niet langer neurotransmitters van de ene naar de andere neuron vuren. Haar hersenen zijn echter bewoond door een zeer kleine en zeer snelle demon, die ingrijpt elke keer als een neuron op het punt zou staan neurotransmitters naar een andere neuron te sturen. Wat de demon veroorzaakt is functioneel hetzelfde als het echt afvuren van neurotransmitters tussen neuronen en gebeurt met dezelfde snelheid. De hersenen van de vrouw werken dus evengoed als gezonde hersenen.

De vraag van Haugeland aan Searle is, kunnen we van deze vrouw zeggen dat ze denkt, of is wat zij doet slechts signaleren op een kunstmatige manier? In het geval van Haugelands demon kiest Searle echter wel de kant van de spreker.

Zijn reden hiervoor is dat de vrouw namelijk wel de juiste causale vermogens heeft, ook al moeten ze een handje geholpen worden door de demon. Hofstadter stelt dat als de systeemreactie in dit geval overtuigend is, het ook overtuigend zou moeten zijn bij het gedachte-experiment van de Chinese kamer.

De vijf instelbare knoppen zijn echter niet de enige parameters, er is nog een hele belangrijke, zegt Hofstadter. Dat is die van het gezichtspunt van waaruit we naar het experiment kijken. Searle kijkt vanuit de persoon in de kamer en ziet daarmee volgens Hofstadter een belangrijk ander perspectief uit het oog. Namelijk dat van de persoon buiten de kamer, de spreker. De systeemreactie komt volgens Hofstadter neer op een verschil in perspectief. Voorstanders van dit bezwaar zouden zeggen dat we vanuit het perspectief van de spreker moeten kijken, niet vanuit het perspectief van de persoon in de kamer.

2.4 Zenon Pylyshyn

Searle beargumenteert in zijn artikel dat alleen machines gemaakt van de juiste materie, hersenen gemaakt van hersencellen, intentionaliteit produceren. In zijn commentaar "The "causal power" of machines"[28] bestrijdt Zenon W. Pylyshyn dit standpunt. Pylyshyn verwoordt het probleem op een andere manier dan Searle. Waar de laatste het heeft over het begrijpen van een verhaal en intentionaliteit, gebruikt Pylyshyn de term 'referentie' of 'verwijzing'. De vraag wordt dan: welke materie heeft het vermogen te verwijzen? Machines ontbreekt het volgens Searle aan intentionaliteit en het vermogen om te verwijzen, omdat ze andere causale vermogens hebben dan ons. Systemen die functioneel identiek zijn hoeven dus niet dezelfde causale vermogens te hebben.

Voor Searle hangt intentionaliteit sterk samen met specifieke materiële eigenschappen, het wordt er letterlijk door veroorzaakt. Stel dat de cel voor cel alle hersencellen van een gezond persoon worden vervangen door chips, met dezelfde input- en outputfunctie als de cel die ze vervangen. De hersenen gemaakt van computerchips zijn functioneel identiek aan de hersenen gemaakt van

[28] Zenon W. Pylyshyn, "The 'Causal Power' of Machines," *Behavioral and Brain Sciences* 3, no. 03 (1980): 442. doi:10.1017/S0140525X0000594X.

hersencellen. Omdat volgens Searle alleen hersenen de juiste causale vermogens hebben om intentionaliteit te veroorzaken, verliezen de hersenen als er genoeg hersencellen zijn vervangen door chips, alle intentionaliteit. Chips hebben nou eenmaal niet de juiste causale vermogens. Voor de buitenstaander verandert er echter niets. Dit staat in contrast met de reactie van Searle op het verhaal van de vrouw met een homunculus in haar hoofd uit het argument van Hofstadter. Searle brengt in de *author's reply* hier tegenin dat het een empirische vraag is of chips intentionaliteit bevatten.[29] De chips moeten dan wel de juiste causale vermogens als menselijke hersenen bevatten.

Volgens Pylyshyn is het duidelijk dat het niet uitmaakt waar een machine van gemaakt is om intentionaliteit te produceren. Op elk niveau kunnen we het materiaal van de hersenen namaken, of het nou het niveau van cellen, neuronen, protoplasma, moleculen, atomen of elementaire deeltjes is. In traditie met de systeemreactie zegt Pylyshyn dat het niet de cellen zijn waar de intentionaliteit zich bevindt, evengoed dat we die niet kunnen vinden in waterpijpen, computerhandelingen of een homunculus in de Chinese kamer. Bovendien weerlegt Searle zijn eigen standpunt in zijn verweer tegen de systeemreactie door te beweren dat de Engelse demon alles uit het hoofd leert en alsnog geen Chinees begrijpt. Hij is gemaakt van de juiste materie, maar begrijpt nog steeds niets. Het kan dus niet aan de materie liggen.

Vervolgens verdedigt Pylyshyn het functionalisme dat Searle aanvalt. Om een antwoord te kunnen geven de vraag wat refereert, moeten we eerst weten waardoor we überhaupt kunnen zeggen dat een symbool refereert. De functionalistische aanpak voor het beantwoorden van deze vraag is door te kijken naar de functionele rol die een symbool speelt in het algehele gedrag van het systeem. Deze manier is niet perfect. Het functionalistische antwoord kan tekort schieten als het niet de vervolgstap neemt om te specificeren wat aan het systeem de toeschrijving van betekenis aan functionele toestanden rechtvaardigt. Het systeem gedraagt zich op een bepaalde manier omdat bepaalde uitdrukkingen een

[29] Searle, "Intrinsic Intentionality," 453.

bepaalde semantische interpretatie hebben, met andere woorden omdat ze bepaalde dingen vertegenwoordigen.

In zijn commentaar legt Pylyshyn uit dat er een probleem is met deze aanpak van het probleem waar Searle op inspringt. De interpretatie is extrinsiek. Ook zonder de interpretatie zou het systeem zich gedragen zoals het zich gedraagt. Searle trekt hieruit de conclusie dat functionalistische theoretici betekenis toeschrijven aan de output, niet dat het systeem er betekenis aan toeschrijft. Het systeem weet niet waar zijn gedrag naar verwijst, de theoretici weten dat. Omdat het systeem niet weet waar zijn gedrag naar verwijst, kan er niet worden gezegd dat het systeem zich op een bepaalde manier gedraagt, door wat het vertegenwoordigt. Mensen gedragen zich wel op basis van de inhoud van onze gedachten.

De stappen die Searle hierboven neemt en de conclusie die hij vervolgens trekt zijn echter een *non sequitur* volgens Pylyshyn. Alleen omdat de theoretici de interpretatie van de output verschaffen, betekent dit niet dat die interpretatie van de theoretici is en niet van het systeem. Het gaat namelijk om de redenen die de theoreticus heeft om een bepaalde interpretatie toe te schrijven. De vraag of de interpretatie in het hoofd van de programmeur of in de machine zit is niet de goede vraag. De relevante vraag is wat de semantische interpretatie van functionele toestanden bepaalt. Theoretici hebben veel vrijheid in het toeschrijven van een semantische interpretatie aan de toestand van een systeem.

Volgens Pylyshyn kunnen we niet met zekerheid zeggen wat het is dat ons in staat stelt te zeggen dat mensen verwijzen. Het argument dat Searle gebruikt tegen de intentionaliteit van computers is volgens hem dus een *argumentum ad ignorantiam*. Als conclusie spoort hij ons aan om nederig te zijn. We moeten toegeven dat we weinig weten over het toeschrijven van intentionaliteit en andere vermogens van computers, ook al weten we wel wat er vanbinnen gebeurt.

In de *author's reply* zegt Searle dat uit het verhaal van Pylyshyn over verwijzing, blijkt dat deze laatste het onderscheid tussen intrinsieke en afgeleide

intentionaliteit niet begrijpt.[30] Dat theoretici veel vrijheid hebben in het toeschrijven van intentionaliteit van computers doet er niet toe. Het belangrijkste is de vraag of deze intentionaliteit intrinsiek is, of afgeleid. Aangezien Searle al voldoende bewezen acht dat alleen mensen intrinsieke intentionaliteit bezitten, gaat hij niet verder in op de functionalistische aanpak van Pylyshyn.

2.5 Daniel Dennett

In de *open peer commentary* van Daniel Dennett op het artikel van Searle, haalt Dennett uit naar Searle en wat hij beweert in "Minds, brains and programs".[31] Dennett heeft eerst kritiek op de vorm van het argument van Searle, later op de inhoud van het argument. Hij zegt namelijk dat Searle de lezer misleidt, zodat deze zijn standpunt accepteert en dat dit standpunt onjuist is. In 2013, in zijn boek *Intuition Pumps and other Tools for Thinking* heeft Daniel Dennett opnieuw kritiek op de Chinese kamer.[32]

Op het niveau van de vorm van het argument zegt Dennett dat Searle sofistiek gebruikt om zijn punt duidelijk te maken, waarmee hij de lezer misleidt. Searle maakt gebruik van een zogenaamde 'intuïtiepomp'. Dit is een middel om intuïties uit te dagen door een variatie op een gedachte-experiment te geven. Deze term is verzonnen door Dennett zelf. Hij vindt dit 'pedagogische hulpmiddel'[33] een fijne manier om de lezer te overtuigen van zijn gelijk. Dit hulpmiddel kan nuttig zijn om de verbeelding van de lezer aan te spreken.

In het geval van de Chinese kamer is het echter een defecte intuïtiepomp die de lezer misleidt en probeert zijn verbeeldingsvermogen uit te schakelen. Het probleem begint als Searle een Schank-computer beschrijft, omdat deze alleen linguïstische input en output kan verwerken. Zo'n computer zal altijd blind blijven

[30] Ibid.
[31] Daniel C. Dennett, "The Milk of Human Intentionality," *Behavioral and Brain Sciences* 3, no. 03 (1980): 428-430. doi:10.1017/S0140525X0000580X.
[32]Daniel C. Dennett, *Intuition Pumps: and Other Tools for Thinking* (New York: W. W. Norton & Company, 2013), 319-329.
[33] Dennett, "The Milk of Human Intentionality," 429.

voor betekenis. Door een computer te beschrijven die niet vergelijkbaar is met de hersenen misleidt Searle de lezer om deze zo van zijn standpunt te overtuigen.

Bovendien laat Searle de lezer een machine voorstellen die veel minder complex is dan de computersimulatie van de Chinese kamer echt zou zijn. Hierdoor ziet Searle de gevolgen van de complexiteit over het hoofd. We weten namelijk niet op welk niveau Searle de handelingen uitvoert. Een computer opereert op verschillende niveaus, die toegang hebben tot verschillende kennis. Als Searle bij wijze van spreken in de kelder van het systeem zit, dan voert hij alleen razendsnelle berekeningen uit. Zit Searle hoger in het systeem, op het niveau van de broncode, dan heeft hij toegang tot opmerkingen over wat hij aan het doen is. In dat geval zou hij hints kunnen lezen die ervoor zorgen dat hij betekenis kan toeschrijven aan de, voor hem, tot dan toe betekenisloze handelingen.

De systeemreactie, zoals ik deze in het eerste hoofdstuk heb besproken, is gestoeld op de kritiek dat Searle twee niveaus van verklaring en toeschrijving door elkaar haalt. Dit zijn het niveau van het geheel en het niveau van de hersenen. Volgens Dennett kun je zeggen: ik begrijp Chinees, mijn hersenen niet. Het gedeelte van mijn hersenen dat begrijpt kan niet geïsoleerd worden. We kunnen niet een plek in de hersenen aanwijzen waar begrip zich bevindt. Voor Searle is dit onbegrijpelijk. Voor hem is het prima mogelijk om te zeggen dat zijn hersenen Chinees begrijpen, evenals hij kan zeggen dat zijn spijsverteringskanaal zijn pizza verteert.[34]

Het grote verschil tussen computers en hersenen is volgens Searle dat de hersenen intentionaliteit voortbrengen en een computer niet. Volgens Dennett en vele KI-experts, zo zegt hij, is het product van de hersenen juist controle, controle over de relaties tussen zintuiglijke input en output in de vorm van gedrag. De hersenen hebben controle over het lichaam doordat ze intentionaliteit produceren. Een computer kan controle hebben zonder intentionaliteit. Controle is volgens Dennett hierdoor het hoofdproduct van de hersenen. Om dit standpunt te

[34] Searle, “Intrinsic Intentionality,” 451.

ondersteunen zegt hij dat een brein zonder controle, maar met intentionaliteit niet kan functioneren. Andersom kan dit wel.

Searle doet het verhaal over controle over input en output af als overblijfsel van het behaviorisme. Een computer kan misschien wel de turingtest halen en ons overtuigen dat het intentionaliteit bevat. Zodra we echter weten dat we met een computer te maken hebben, weten we dat het geen intentionaliteit kan bezitten.[35] Het hebben van de juiste input-outputrelaties is voor Searle slechts een symptoom van intentionaliteit, geen definitief bewijs ervoor. Het gaat voor hem erom wat er aan de binnenkant aan de hand is.

Wat moet er dan precies aan de binnenkant aan de hand zijn? Dat zijn volgens Searle de causale eigenschappen die intern zijn aan de werking van de hersenen. Deze causale eigenschappen zijn alleen intern aan de hersenen, niet aan een computer. We moeten dus op zoek naar de juiste causale eigenschappen. Hij bedoelt hiermee niet de eigenschappen die de activiteiten van de hersenen met de dingen in de buitenwereld verbinden. Wat hij er wel mee bedoelt wordt niet precies duidelijk.

Hoe kunnen we echter weten dat een computer deze juiste eigenschappen mist, als we alleen weten dat het een formeel programma is? Dennett zegt in zijn commentaar dat Searle toegeeft dat we de hersenen kunnen beschrijven in termen van een formeel programma. We kunnen namelijk alleen intentionaliteit toeschrijven op basis van wat er vanbinnen gebeurt. Zodra we weten dat er geen geest aanwezig is, maar dat we te maken hebben met de implementatie van een formeel programma, kunnen we geen intentionaliteit meer toeschrijven.

De titel van Dennetts artikel, “The Milk of Human Intentionality” verwijst naar een vergelijking die Searle maakt tussen lactatie en intentionaliteit. Searle zegt namelijk dat een computersimulatie van het lactatieproces evenmin melk produceert als een simulatie van de hersenen intentionaliteit produceert.[36] Dennett zegt dat niets alleen de implementatie van een formeel programma is. Een werkende computer scheidt namelijk tijdens zijn handelingen warmte uit, waarom

[35] Searle, “Minds, Brains and Programs,” 421.
[36] Ibid., 424.

niet ook een beetje intentionaliteit?[37] Searle is het hier niet mee eens. Hij bedoelde met zijn vergelijking tussen lactatie en de hersenen niet dat intentionaliteit een soort melk is die door het lichaam wordt uitgescheiden, maar dat intentionaliteit alleen door en in de structuur van de hersenen gerealiseerd kan worden.[38]

Dennett denkt dat het feit Searle zo bezig is met de interne eigenschappen van controlesystemen voortkomt uit zijn poging om het interne standpunt van een wezen met bewustzijn te pakken te krijgen. Searle ziet niet in hoe een computer van zichzelf kan denken dat het bewust is. Dit komt omdat hij te diep graaft. We kijken ook niet naar de synapsen in de hersenen op zoek naar bewustzijn. Hij zit op het verkeerde beschrijvingsniveau. De systeemreactie komt volgens Dennett simpelweg neer op de bewering dat Searle bewustzijn en intentionaliteit op de verkeerde plek zoekt.

In 2013 voegde Dennett aan deze eerste verdediging van de systeemreactie een extra argument toe.[39] Searle vindt het gênant dat hij überhaupt de systeemreactie moet weerleggen, het lijkt voor hem zo overduidelijk een implausibele theorie.[40] Zoals ik al eerder heb genoemd zouden volgens Searle alleen aanhangers van een ideologie de systeemreactie steunen. Dennett zegt dat deze ideologie de kern van de computerwetenschappen is. Wat Searle zo implausibel vindt, zo zegt Dennett, is het fundamentele inzicht van Alan Turing dat het gaat om de software. De kracht van een computer zit in het systeem, niet in de onderliggende hardware.

Intentionele houding

In zijn essay “Intentional Systems” in het boek *Brainstorms* legt Daniel Dennett drie stances, of houdingen, uit waarop we het gedrag van bijvoorbeeld een computer kunnen verklaren.[41] Deze drie houdingen zijn de ontwerphouding, de

[37] Dennett, “The Milk of Human Intentionaliy,” 430.
[38] Searle, “Intrinsic Intentionality,” 451.
[39] Dennett. *Intuition Pumps*, 319-329.
[40] Searle, “Minds, Brains and Programs,” 419.
[41] Dennett, *Brainstorms*, 3-22.

fysieke houding en als laatste, en voor ons de meest relevante, de intentionele houding.

De eerste houding is de ontwerphouding. Als we precies weten hoe een computer ontworpen is, kunnen we op basis van het ontwerp het gedrag verklaren, mits de computer doet wat geprogrammeerd is. Deze houding is nuttig bij het verklaren van mechanische objecten, maar ook lucifers en struiken. Het gedrag van computers kan worden voorspeld door middel van de ontwerphouding, maar het ontwerp van een computer is te complex om dit met gemak en enige snelheid te doen.

De tweede houding is de fysieke houding. Deze houding gaat uit van de fysieke samenstelling van een object en de natuurwetten die van toepassing zijn. Met deze houding kunnen we niet geprogrammeerde storingen van een systeem voorspellen. Ook de fysieke houding kan worden gebruikt om het gedrag van een computer te voorspellen, maar duurt wederom te lang vanwege de complexiteit van computers. Het is beter geschikt voor het maken van voorspellingen over defecte machines, waarbij de oorzaak van het defect makkelijk te vinden is, zoals een machine met de stekker uit het stopcontact.

De beste manier om het gedrag van een computer te voorspellen is ervan uit te gaan dat de computer functioneert zoals deze is ontworpen, dat het ontwerp optimaal is en dat de computer de meest rationele optie zal kiezen. Dat wil zeggen, de computer behandelen als een intelligent mens. Dit is de intentionele houding, waarbij men naar de computer kijkt alsof het een intentioneel systeem is. Dennett definieert een intentioneel systeem als een systeem wiens gedrag we kunnen verklaren en voorspellen door middel van intentionele toestanden, zoals overtuigingen en verlangens, aan het systeem toe te schrijven. Iets is alleen een intentioneel systeem in verhouding tot de strategieën van degene die het gedrag van dit systeem probeert te verklaren en te voorspellen. De intentionele houding is een voor de hand liggende houding ten opzichte van computers, maar ook ten opzichte van dieren en andere mensen. We zijn van nature geneigd om anderen te beschouwen als rationeel en intentioneel.

De intentionele houding begint met het veronderstellen van rationaliteit. We gaan ervan uit dat een computer informatie bezit en doelen heeft die het wil bereiken. Als we ervan uitgaan dat een computer rationeel is, is het een kleine stap om te zeggen dat de informatie die de computer bezit ‘overtuigingen’ zijn en de doelen ‘verlangens’. We voorspellen zo het gedrag van een computer als antwoord op de vraag wat het meest rationele gedrag zou zijn op basis van zijn overtuigingen en wensen.

Het bezitten van informatie is intentioneel. Het is geen saaie en onschuldige notie van opslag, het is epistemisch bezit. Vergelijk het met het bezitten van een encyclopedie en het bezitten van de kennis in een encyclopedie. Een computer moet op deze manier gezegd worden informatie te bezitten om deze te kunnen gebruiken om zo beslissingen te maken. Deze manier van praten over informatie lijkt op het onderscheid dat Searle maakt tussen de informatieverwerking die een mens doet en die een computer doet. De informatieverwerking van een mens is intentioneel, wij weten immers waar de informatie die we verwerken over gaat. Dennett zegt dat we bij de intentionele houding moeten aannemen dat de computer net als de mens weet waar de informatie over gaat.

Deze houding zegt niets over of de computer werkelijk interne toestanden heeft, maar dat we deze aan de computer toeschrijven. Dit is volgens Dennett de beste manier om met computers om te gaan. Computers zijn namelijk te ingewikkeld voor de ontwerphouding of de fysieke houding. De intentionele houding is een pragmatisch besluit, niet intrinsiek goed of fout. Het maakt niet uit of een computer werkelijk intentionele toestanden heeft, we schrijven ze toe om een pragmatische reden. Het maakt ook niet uit waar een computer van gemaakt is. Dit alles zegt niet dat een computer een adequaat model van de geest of menselijke intelligentie is, of dat het een simulatie ervan is. Wat Dennett wil zeggen met de intentionele houding is dat het handig kan zijn om een computer, een complex, gestructureerd en puur fysiek systeem te behandelen alsof het rationeel is.

Dennett erkent dat er een belangrijke tekortkoming van de intentionele houding is, namelijk dat het rationaliteit en intelligentie veronderstelt, maar het niet uitlegt. De houding werkt niet meer zodra we teveel gaten in de rationaliteit van een computer ontdekken. Wanneer dit gebeurt moeten we de intentionele houding verhouden en overstappen op de ontwerphouding. Dit is ook het uiteindelijke doel van KI volgens Dennett, namelijk de intelligentie van mens en machine uit te leggen in termen van het ontwerp.

Dat de intentionele houding niet perfect is, maakt deze niet minder nuttig. Het is makkelijker om te beslissen of een machine intentionaliteit bevat dan de vraag of de machine echt kan denken. Een mens is een intentioneel systeem, de rest volgt hieruit. KI werkt vanuit een intentioneel gekarakteriseerd probleem naar een oplossing vanuit de ontwerphouding. De eenvoud van het behandelen van een computer als intentioneel systeem maakt het ideaal als bron voor orde en organisatie in filosofische analyses van mentale concepten.

Representatie en het nut van KI

In *Brainstorms* wijdt Dennett zijn eigen theorie over interne representaties uit.[42] Hij legt dit uit in de wijdere context van de bijdragen van KI aan het abstracte onderzoek naar intelligentie en kennis en het probleem dat ontstaat als psychologie wordt uitgelegd in termen van representatie.

Dennett legt dit probleem als volgt uit. De enige theorie die de complexiteit van menselijke activiteit kan verklaren moet interne representaties poneren. Dit is wat Searle doet in zijn definitie van intentionaliteit, zoals we hebben gezien in het eerste hoofdstuk. Dit is de eerste premisse, die makkelijk geaccepteerd werd, met uitzondering van radicale behavioristen.

De tweede premisse stelt dat niets intrinsiek een representatie is, iets is alleen een representatie ten opzichte van iemand. Elke interpretatie heeft, met andere woorden, iemand nodig die de representatie gebruikt of interpreteert. Deze gebruiker moet intentionele eigenschappen hebben en overtuigingen en doelen,

[42] Dennett, *Brainstorms*, 119-25.

zodat het de representaties kan gebruiken om hem te helpen in het bereiken van deze doelen. Zo'n gebruiker is een homunculus.

De combinatie van deze twee premissen geeft een problematische conclusie. Psychologie zonder homunculus is onmogelijk, maar psychologie met homunculi is gedoemd tot een cirkelredenering of een oneindige regressie. De conclusie is dus dat psychologie onmogelijk is. Dit probleem noemt Dennett Humes probleem. Dennett noemt dit probleem zo omdat volgens hem Hume de eerste is die met dit probleem worstelde. Om dit probleem op te lossen moeten representaties in staat zijn zichzelf 'te begrijpen'. Searle ontkent zoals we hebben gezien in hoofdstuk 1 de tweede premisse, waarmee hij Humes probleem omzeilt.

De topdown aanpak van kunstmatige intelligentie geeft volgens Dennett de oplossing voor dit probleem. Er is een oneindige regressie van homunculi als er minstens een homunculus is die de representaties interpreteert. We vinden de oplossing voor dit probleem wanneer we de regressie bekijken als een computer. Het eerste niveau is het intentionele systeem. Het tweede niveau zijn intentionele subsystemen, de homunculi. De verdere niveaus bestaan uit progressief simpelere homunculi. Het laagste niveau is een homunculus die een simpele ja/nee-taak uitvoert. Dit niveau kan worden vervangen door een machine. De regressie is hierdoor niet langer oneindig.

Dennett noemt nog een tweede probleem dat filosofie samen met KI kan oplossen, namelijk het frame-probleem.[43] Dit is een abstract epistemologisch probleem dat ontdekt is door KI gedachte-experimenten. Wanneer een cognitief wezen, een entiteit met overtuigingen over de wereld, een handeling uitvoert, verandert de wereld. Hierdoor moeten de overtuigingen van het wezen over de wereld worden herzien en bijgewerkt. Hoe gebeurt dit? We kunnen niet alle veranderingen opmerken, dus kunnen we niet volledig op onze zintuiglijke informatie vertrouwen om onze overtuigingen te herzien. We moeten interne manier hebben om onze overtuigingen bij te werken, zodat ons interne model van de wereld weer grofweg overeenkomt met de wereld.

[43] Dennett, *Brainstorms*, 125-26.

Traditioneel werd door filosofen veronderstelt dat iemands overtuigingen een set van proposities zijn en dat redeneren deductie van leden van deze set is. Hier beginnen volgens Dennett de moeilijkheden. Het is gebleken dat systemen die op zulke processen werken, overspoeld worden door een veelheid van mogelijke combinaties in de poging de overtuigingen bij te werken. Om de niet te ontkennen capaciteit van mensen om hun overtuigingen in overeenstemming te brengen met de wereld uit te kunnen leggen, moeten we onze hele opvatting van overtuigingen en redeneren radicaal herzien.

Deze twee filosofische problemen laten zien dat het project van sterke KI ons wel degelijk iets kan leren over de menselijke psychologie en cognitieve toestanden, in tegenstelling tot wat Searle denkt.

3. Inventarisatie van de argumenten

Nu we de argumenten voor en tegen het gedachte-experiment van de Chinese kamer op een rij hebben, is het zaak om de argumenten tegen elkaar af te wegen om zo te bepalen welke van de twee kampen gelijk heeft. Ik zal hieronder de argumenten tegen Searle op thema indelen. Als eerste zal ik de kritiek op het gebrek aan onderscheid tussen functionalisme en behaviorisme behandelen. Vervolgens de kritiek op de vereenvoudiging van wat er aan de hand is door Searle. Daarna zal ik de twee belangrijkste kritiekpunten behandelen, namelijk de onduidelijkheden ten opzichte van 'de juiste materie' en de kritiek met betrekking tot perspectief.

3.1 Functionalisme en behaviorisme

Een van de problemen die Rey heeft met het argument van Searle is dat hij het functionalisme en het behaviorisme niet goed van elkaar onderscheidt. Het functionalisme lijkt op het behaviorisme, maar er is een belangrijk verschil dat van groot belang is in deze discussie. Door het functionalisme niet goed te onderscheiden van het behaviorisme, dat tijdens het schrijven van de tekst van Searle uit de mode was geraakt, misleidt hij de lezer met een vooringenomen houding jegens het functionalisme. Behaviorisme wil gedrag verklaren en voorspellen op basis van actueel en dispositioneel gedrag. In een poging om het lichaam-geestdualisme op te lossen, zien behavioristen mentale toestanden als disposities tot een bepaald gedrag. Ook in zijn behandeling van de combinatiereactie maakt Searle geen onderscheid tussen het behaviorisme en het functionalisme dat hij wil aanvallen.

Het functionalisme lijkt op het behaviorisme omdat het een grote nadruk legt op de input en de output. De functionalist ontkent mentale toestanden echter niet. De in- en output moet worden bemiddeld door de juiste interne toestanden. Wat er binnen gebeurt is voor de behaviorist niet belangrijk, maar voor de functionalist en voor Searle is dit wel van belang. Als wat er binnen een computer

gebeurt functioneel equivalent is aan wat er in een mens gebeurt, dan zal de functionalist zeggen dat de computer, in het geval van de Chinese kamer, Chinees begrijpt.

Het verwarren van het functionalisme met het behaviorisme heeft grote gevolgen voor kritiek op het functionalisme. Een behaviorist zal eerder zeggen dat een computer intelligent is, namelijk als deze zich intelligent gedraagt. Searle legt veel nadruk op zijn overtuiging dat het hebben van de juiste input, het juiste programma en de juiste output niet genoeg is voor intentionaliteit. Een behaviorist zou het niet met deze overtuiging eens zijn, omdat deze zou zeggen dat we kunnen weten of iemand Chinees begrijpt door naar het gedrag te kijken. Een functionalist zou het met Searle eens zijn. De juiste input en output, bemiddeld door het juiste programma is niet genoeg voor intentionaliteit, wat er gebeurt in de computer moet functioneel identiek zijn aan wat er gebeurt in een mens. Dat is niet het geval in het Chinese kamer gedachte-experiment. We kunnen dus stellen dat de kritiek van Searle zijn doel mist.

3.2 Vereenvoudiging

Rey verwijt Searle ook dat hij het begrijpen van een taal simpeler voorstelt dan het in werkelijkheid is. De persoon in de kamer volgt in de systeemreactie alleen de regels van de taal, maar wat nodig is voor begrip zijn programma's die te maken hebben met de relatie tussen de mentale wereld en de buitenwereld. Iemand begrijpt pas een taal als hij snapt dat de geuite tekens betrekking hebben op bepaalde waarnemingen, geloven en wensen.

Dit laatste is echter precies het punt van Searle. Iemand begrijpt pas Chinees als zijn mentale toestanden intentioneel zijn. Volgens Searle is het draaien van een programma niet genoeg voor intentionaliteit, maar Rey lijkt hier te zeggen dat we gewoon het intentionaliteitsprogramma moeten draaien. Het is waar dat Searle het gedachte-experiment te simpel voorstelt. Anderzijds stelt Rey op zijn beurt intentionaliteit te simpel voor en zijn kritiek is daarom niet toereikend om Searles nadruk op het belang van intentionaliteit te ontkennen.

Ook Hofstadter is van mening dat Searle de zaken verkeerd voorstelt aan de lezer door ze vereenvoudigt weer te geven. Het is namelijk een flinke stap van het beantwoorden van vragen over een Chinees restaurant naar het slagen voor de turingtest. Bovendien zijn beide scenario's ongeloofwaardig. Een mens zou veel te lang doen over deze processen om aan te worden gezien voor een Chineessprekend persoon.

Het vereenvoudigen van ingewikkelde problematiek, zoals in het geval van het argument van de Chinese kamer is bedoeld om het makkelijker te maken voor de lezer om te begrijpen wat er aan de hand is. Door het begrijpen van taal eenvoudiger voor te stellen dan het is en door de complexiteit van de handelingen van de persoon in de Chinese kamer simpeler te beschrijven, misleidt Searle de lezer door een verkeerde stand van zaken voor te spiegelen. Zowel Rey als Hofstadter hebben hier terecht kritiek op.

3.3 De juiste materie?

Searle legt veel nadruk op het idee dat alleen onze hersenen de juiste causale eigenschappen hebben die nodig zijn voor intentionaliteit. Zijn vasthoudendheid zorgt echter voor onduidelijkheid. Zoals we hebben gezien in het gedachte-experiment van Haugeland, aangehaald door Hofstadter, werd de functie van neuronen overgenomen door een demon. Wat de demon veroorzaakt is functioneel equivalent aan wat er gebeurt in gezonde hersenen. De door de demon gestuurde hersenen werken evengoed als gezonde hersenen. Omdat deze persoon nog wel de juiste causale vermogens bevat, is er volgens Searle sprake van intentionaliteit.

In deze aangepaste vorm van de systeemreactie is Searle het wel eens met zijn critici. De juiste materie is voor Searle zo belangrijk dat het voor hem niet uitmaakt of deze wordt gestuurd door een demon wanneer de hersenen hun functie niet zelf kunnen uitvoeren. Hofstadter zegt mijns inziens terecht dat de systeemreactie ook in de reguliere vorm overtuigend is. Bovendien zijn de intuïties van Searle over de rol van de hersenen bij het veroorzaken van

intentionaliteit geen goede reden om het ontstaan van intentionaliteit uit een ander systeem af te wijzen.

Een ander gedachte-experiment waarbij de functie van de hersenen wordt overgenomen, vinden we bij Pylyshyn. In zijn versie wordt dat gedaan door computerchips, niet door een demon of persoon in een kamer. Als er genoeg hersencellen worden vervangen door de chips, verliest de persoon alle intentionaliteit, omdat de juiste materie wordt vervangen door de computerchips. Het maakt voor Searle niets uit dat een buitenstaander niets zou veranderen. Het is onduidelijk wanneer precies alle intentionaliteit verloren zou gaan.

Searle geeft als kritiek op Pylyshyn dat het een empirische vraag is of computerchips intentionaliteit bevatten. Dit staat in groot contrast met de nadruk die Searle in zijn argument van de Chinese kamer legt op causale eigenschappen die alleen kunnen voortkomen uit de juiste materie, de hersenen. Deze nadruk op de juiste materie ondermijnt bovendien de kritiek van Searle op de systeemreactie. Als de Engelstalige Searle namelijk alles uit het hoofd leert, maar toch geen Chinees begrijpt, kan dat niet aan de juiste materie liggen.

In *Intentionality: An Essay in the Philosophy of Mind* geeft Searle ons een mogelijke hint wat volgens hem wel de juiste materie is.[44] We schrijven intentionaliteit toe aan honden omdat we zo hun gedrag kunnen verklaren en omdat ze van dezelfde materie zijn gemaakt als mensen. Hij noemt in dit boek de ogen, huid en oren van een hond als aanwijzing dat dieren intentionaliteit bevatten. Door deze zinsnede lijkt hij de robotreactie te bevestigen, namelijk dat als we een robot kunnen laten waarnemen, kunnen we intentionaliteit toeschrijven.

De vasthoudendheid van Searle aan het belang van de juiste materie is naar mijn mening niet te rechtvaardigen. Dit omdat Searle zelf onduidelijk is over wat de juiste materie is en hij geen goede onderbouwing heeft waarom alleen de hersenen intentionaliteit bevatten. Bovendien stelt hij dat het een empirische vraag is welke materie intentionaliteit bevat. Deze opmerking staat in scherp contrast

[44] Searle, *Intentionality*, 5.

met zijn stellige overtuiging dat alleen mensen intentionaliteit bevatten omdat alleen de hersenen uit de juiste materie zijn gemaakt.

3.4 Perspectief en intentionaliteit

Intentionaliteit en perspectief spelen samen een sleutelrol in het argument van Searle. Zijn hoofdpunt is namelijk dat een machine geen intrinsieke, maar alleen afgeleide intentionaliteit bevat. Het perspectief van waaruit dit wordt beoordeeld doet er niet toe. Gezien vanuit het perspectief van zowel de machine als dat van de buitenstaander heeft de machine geen intrinsieke intentionaliteit. Dit is precies waar het Searle om gaat. In de tekst “Minds, Brains and Programs” heeft Searle het echter nooit expliciet over dit onderscheid of het belang ervan voor zijn argument van de Chinese kamer. Hij zegt alleen dat we niet zomaar intentionaliteit aan van alles kunnen toeschrijven, iets moet echt intentionaliteit bezitten. Hofstadter en Pylyshyn zeggen dat je prima intentionaliteit kan toeschrijven aan een machine en dat Searle op de verkeerde plek naar intentionaliteit zoekt.

De systeemreactie kunnen we volgens Hofstadter samenvatten als een verwarring over perspectief. Searle zegt dat we geen intentionaliteit en dus begrip zien bij de persoon in de kamer, ook al heeft hij alle regels uit het hoofd geleerd. We moeten echter niet kijken vanuit het perspectief van Searle, maar vanuit het perspectief van de persoon buiten de kamer. De intentionaliteit die we als waarnemer zien, is echter afgeleide intentionaliteit. Dit argument van Hofstadter toont aan dat een machine wel afgeleide intentionaliteit heeft. Dit zou Searle echter zelf ook niet ontkennen.

Ook Pylyshyn heeft het over het perspectief. Wanneer hij het heeft over de interpretatie van het gedrag van een computer door theoretici, bekijken we de computer vanuit het perspectief van de buitenstaander. Dit maakt voor Pylyshyn deze interpretatie niet minder waard dan het perspectief van de computer zelf, omdat de waarde van de interpretatie van de theoreticus afhangt van zijn redenen voor deze interpretatie. Voor Searle kunnen de redenen voor een interpretatie van

het gedrag van een computer nog zo goed zijn, dit maakt niet dat de computer intrinsieke intentionaliteit bezit. Pylyshyn zelf maakt geen onderscheid tussen de twee soorten intentionaliteit.

Daniel Dennett lost het probleem van intentionaliteit in computers op door zich niet af te vragen of computers intentionaliteit bevatten, maar door het vanuit een pragmatische oogpunt te bekijken. Voor hem is het praten over computers als intentionele systemen een pragmatische houding, omdat het als beste het gedrag van zulke complexe machines kan verklaren en voorspellen. Searle ziet een valkuil in het toeschrijven van intentionaliteit aan computers. Als we geavanceerde computers behandelen als intentionele systemen, is de stap naar bijvoorbeeld een thermostaat snel gemaakt. Searles kritiek gaat echter niet op, omdat het voor minder geavanceerde technologieën als een thermostaat praktisch meer zin heeft om deze te benaderen vanuit de ontwerphouding.

Als KI-onderzoek menselijke cognitieve toestanden wil nabootsen, dan moet een computer intrinsieke intentionaliteit bevatten. Dit maakt het onderscheid belangrijk, maar Searle heeft niet duidelijk aangetoond waarom computers geen intrinsieke intentionaliteit zouden kunnen bezitten. De conclusie van Searle dat het project van sterke KI onmogelijk is, is dus niet eenduidig bewezen.

Conclusie

Door de toenemende ontwikkelingen in kunstmatige intelligentie en in de neurowetenschap, wordt de vraag over echt slimme computer steeds relevanter en belangrijker. KI-wetenschappers ontwerpen aanzienlijk menselijkere machines, waardoor we moeten herzien welke eigenschappen specifiek menselijk zijn. Door deze toenemende relevantie is het van belang dat filosofen mee blijven denken in het debat over KI.

Searle heeft de vraag of machines kunnen denken uitgedrukt in de vraag of machines intrinsieke intentionaliteit kunnen bezitten. Zijn antwoord hierop was ‘nee’. Hij is echter met name wat betreft de materie en intentionaliteit niet overtuigend dat computers geen intrinsieke intentionaliteit kunnen hebben. Het is aan de andere kant ook niet bewezen door de door mij behandelde filosofen dat machines wel intentionaliteit kunnen hebben. Kunnen machines denken? Het ongelijk is niet bewezen, de vraag is nog open.

Mijn conclusie is om een oordeel over intrinsieke intentionaliteit toeschrijven aan machines op te schorten, tot het moment dat het onderzoek in kunstmatige intelligentie zover gevorderd is dat er wel een eenduidig oordeel gevormd kan worden.

Bibliografie

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Philosophy of Mind Series; Philosophy of Mind Series. New York: Oxford University Press, 1996.

Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Hassocks: Harvester Press, 1979.

———. *Intuition Pumps and Other Tools for Thinking*. First edition. ed. New York: W.W. Norton & Company, 2013.

———. "The Milk of Human Intentionality." *Behavioral and Brain Sciences* 3, no. 03 (1980): 428. doi:10.1017/S0140525X0000580X.

Hofstadter, Douglas R. en Daniel C. Dennett. *The Mind's I: Fantasies and Reflections on Self and Soul*. Brighton: Harvester Press, 1981.

Hofstadter, Douglas R. "Reductionism and Religion." *Behavioral and Brain Sciences* 3, no. 03 (1980): 433. doi:10.1017/S0140525X00005847.

Pylyshyn, Zenon W. "The ‘Causal Power’ of Machines." *Behavioral and Brain Sciences* 3, no. 03 (1980): 442. doi:10.1017/S0140525X0000594X.

Levin, Janet. "Functionalism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/win2016/entries/functionalism

Ravenscroft, Ian. "Functionalism." In *Philosophy of Mind: A Beginner's Guide*, 50-63. New York: Oxford University Press Inc., 2005.

Rey, Georges. "What's Really Going on in Searle's “Chinese Room”." *Philosophical Studies : An International Journal for Philosophy in the Analytic Tradition* 50, no. 2 (1986): 169-85. doi:10.1007/BF00354586.

Searle, John R. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1983.

———. "Intrinsic Intentionality." *Behavioral and Brain Sciences* 3, no. 03 (1980): 450. doi:10.1017/S0140525X00006038.

———. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 03 (1980): 417. doi:10.1017/S0140525X00005756.

———. *The Rediscovery of the Mind*. Cambridge, Mass.: The MIT Press, 1992.

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433-60.

Verhagen, Laurens. "Slimme computers: kunnen ze straks ook kunst maken?" *Volkskrant*, 15-07-2017.