Filosofie van Kunstmatige Intelligentie, 2008
0440949 Andreas van Cranenburgh
Sun Sep 21 13:00:32 CEST 2008

1 De Chinese kamer

De Chinese kamer is een seminaal gedachtenexperiment van John Searle (1980). Hij vraagt ons een kamer voor te stellen waarin een monolinguistisch, Engelssprekend proefpersoon zich bevindt, samen met een stapel papieren met herschrijfregels voor het beantwoorden van vragen in het Chinees. De proefpersoon begrijpt niets van Chinees maar is wel in staat de vorm van de Chinese karakters te herkennen.

Van buiten zal het voor Chinezen lijken alsof de proefpersoon Chinees spreekt, en deze zal dus slagen voor de Turing test. Echter, er zou duidelijk geen sprake zijn van begrip, en dus is er geen intentionaliteit.

In het gedachtenexperiment staan de herschrijfregels symbool voor een computerprogramma dat intelligent zou zijn, en de proefpersoon staat voor de machine die formele instructies uit kan voeren. Searle is van mening hiermee te hebben aangetoond dat *strong AI* een onhoudbare positie is.

1.1 Het onderscheid tussen *strong* en *weak AI*

Searle maakt een onderscheid tussen *strong* en *weak AI*. De eerste beweert dat een digitale computer voorzien van het juiste programma niet te onderscheiden is van een menselijke geest, in gedrag maar ook in werking. De machine zou dan werkelijke over *mental states* beschikken, en wel dezelfde als die van mensen.

Weak AI neemt een voorzichtigere positie in en beweert enkel dat de computer een nuttig gereedschap biedt om modellen van intelligentie op de proef te stellen. Als een programma schijnbaar intelligent gedrag vertoont zal men tevreden zijn, terwijl de werking niet één op één overeen hoeft te komen met die van een menselijke geest.

Besloten in *strong AI* ligt vaak de these dat de menselijke geest zelf een computer is, het zogenaamde computationalisme. Dit zou betekenen dat het brein, en dus ook het verdere lichaam, totaal niet van belang is voor intelligentie. De geest zou bestaan uit een stuk *software*, dat gebruikt maakt van een machine als ware het een substraat. Deze these, dat de geest onafhankelijk is van het lichaam wordt functionalisme genoemd.

Ten laatste is er het dualisme van lichaam en geest dat Searle specifiek tracht te bestrijden; het geloof dat het mentale en het fysieke uit fundamenteel verschillende substanties bestaan.

Theoretisch gezien is dit een zeer zinnig onderscheid. Wanneer de claim wordt gemaakt dat er sprake is van *strong AI*, dan kan men zeer skeptisch zijn, onder andere door het gedachtenexperiment van Searle. Wanneer echter de claim 'slechts' *weak AI* betreft dan is deze niet zo sterk, omdat het hoogstens een simulatie van menselijke vermogens betreft, zonder dat deze noodzakelijkerwijs iets verklaart over de werking van menselijke intelligentie.

In de praktijk kan men toch grote vraagtekens plaatsen bij de gevolgen van het gedachtenexperiment van Searle. Er zijn niet veel onderzoekers binnen de

AI die expliciet geloven in de *strong AI* hypothese[1]. Het is doorgaans niet van belang *hoe* het werkt, maar vooral *of* het werkt. Het nabootsen van menselijke intelligentie is meestal niet het hoofddoel, maar juist een of ander praktisch doel, waarvan dan ook makkelijker het succes te bepalen is.

In feite is het resultaat van Searle dus belangrijker voor het cognitivisme, dat het computationalisme vaak als uitgangspunt neemt, dan voor de kunstmatige intelligentie, waar veelal geen enkele filosofie wordt aangehangen.

1.2 Wat wordt weerlegd?

1.2.1 Programma's kunnen intentionaliteit hebben

Volgens Searle toont zijn gedachtenexperiment aan dat een set formele regels, die wordt toegepast op een set formele symbolen, niet tot gevolg kan hebben dat de betekenis van de symbolen begrepen wordt. Dit begrip is aanwezig noch in de regels, noch in degene of dat wat de regels uitvoert, noch in de combinatie van deze twee, zolang enkel naar de vorm van de invoer wordt gekeken. Syntaxis alleen kan dus geen aanleiding vormen voor semantiek. Intentionaliteit kan niet ontstaan enkel door het uitvoeren van een computerprogramma.

1.2.2 Wat programma's doen kan de werking van de menselijke geest verklaren

Ten gevolge van het afwijzen van de vorige stelling moet ook, *a fortiori*, deze stelling verworpen worden. Als wordt aangenomen dat mensen *wel* beschikken over intentionele vermogens, begrip hebben van de semantiek van symbolen, dan kan de menselijke geest niet uit louter een digitaal computerprogramma bestaan, aangezien deze slechts syntaxis zou kennen, en geen semantiek.

1.3 Kunnen machines denken?

Machines kunnen niet denken, zolang onder denken het hebben van intentionaliteit wordt verstaan en de machine in kwestie geen menselijk brein is. Dit komt omdat een gewone machine (bijvoorbeeld een digitale computer) niet over dezelfde causale vermogens beschikt als het menselijk brein, ongeacht het programma wat op de computer draait. Deze causale vermogens zijn nou juist degene die intentionaliteit mogelijk maken bij mensen.

Als het menselijk brein wordt beschouwd als een ingewikkelde machine, wat zeer wel mogelijk is met een materialistisch wereldbeeld, dan moet het antwoord op de vraag "ja" zijn, want het is duidelijk dat mensen kunnen denken. Volgens het genoemde wereldbeeld mag worden aangenomen dat "het brein de geest veroorzaakt." Als echter het menselijk brein in zijn geheel moet worden nagebouwd om intelligentie na te kunnen bootsten, dan is er geen sprake meer van *kunstmatige* intelligentie, omdat het geen emulatie betreft, maar simulatie. Hiermee heeft de *strong AI* these afgedaan voor Searle.

[1] Ray Kurzweil gebruikt de term *strong AI* zelfs voor iets heel anders, namelijk voor "sterkere intelligentie dan die van mensen"

1.4 Twee tegenwerpingen

1.4.1 The Robot reply (Yale)

Deze tegenwerping beweert dat het toevoegen van zintuigen en actuators om met de buitenwereld te interacteren als gevolg zou hebben dat er wel sprake kan zijn van begrip en *mental states*.

Volgens Searle verandert dit niets aan zijn gedachtenexperiment. Het experiment wordt simpelweg aangepast aan de nieuwe situatie, net zoals een theorie net zo veel kan worden aangepast aan nieuwe empirische data, zonder dat vast is te stellen of de theorie misschien verworpen moet worden. Searle stelt voor dat de invoer en uitvoer verloopt zoals in het originele experiment, met behulp van Chinese symbolen. Nu zullen sommige van deze symbolen echter afkomstig zijn van een camera, of, in het geval van uitvoer, een actuator in beweging brengen. Hiermee is de proefpersoon alsnog gereduceerd tot een manipulator van formele symbolen.

Hier echter vind ik dat Searle zich er te makkelijk vanaf maakt. Door de perceptuele informatie geïnterpreteerd (want symbolisch) en wel te presenteren beschikt de Chinese kamer dus niet over een werkelijk sensori-motor systeem, zoals dat bij mensen (en dieren) aanwezig is. Naar mijn mening kunnen zintuigen subsymbolische informatie presenteren, ruwe data, als het ware. Deze kan dan later al dan niet worden geinterpreteerd; terwijl het ook mogelijk is dat deze data multi-interpretabel is. Verder moeten de effecten van het motor systeem op de wereld direct waarneembaar zijn, opdat het sensori-motor systeem kan reageren op feedback van de buitenwereld, zoals het compenseren voor *a priori* onbekende krachten zoals de zwaartekracht.

Het is natuurlijk mogelijk dat de informatie niet vooraf geïnterpreteerd wordt, dat in plaats van "rode bal" een grote verzameling pixels wordt gepresenteerd. In dit laatste geval zal het systeem echter vrij zijn de zintuigelijke informatie te interpreteren zoals schikt (bijvoorbeeld volgens zelf geleerde patronen), en is er geen *a priori* reden om dit niet semantisch te mogen noemen. Waar het om gaat is dat er meerdere niveaus van beschrijving bestaan, terwijl Searle het probeert over te doen komen alsof alle symbolen gelijk zijn.

Het sensori-motor systeem mag dus *niet* buiten de Chinese kamer worden geplaatst. Doch moet toegegeven worden dat ook een sensori-motor systeem geen voldoende voorwaarde is voor intentionaliteit, aangezien veel dieren ogenschijnlijk en waarschijnlijk niet over intentionaliteit beschikken, maar wel over een sensori-motor systeem. Wel ben ik van mening dat een sensori-motor systeem een noodzakelijke voorwaarde vormt voor een wereldlijk bewustzijn.

1.4.2 The brain simulator reply (Berkeley & MIT)

Een andere tegenwerping op het Chinese kamer gedachtenexperiment komt van Berkeley en MIT, *the brain simulator reply*, ook wel met een parafrase *The Chinese Gym experiment* genoemd.

Stel we maken een model van een Chinees brein als een neuraal netwerk, bijvoorbeeld in de vorm van waterpijpen, waarbij Chinese symbolen als invoer kunnen worden gegeven door ze op een geschikte manier te representeren als *neuron firings*.

Volgens Searle zal noch de operator van het neurale model, noch het neurale model, noch de combinatie van deze twee beschikken over intentionele staten,

oftewel begrip hebben van de betekenis van de verwerkte symbolen.

Hier voert Searle zijn gedachtenexperiment wel erg ver, en blijkt vooral welk een magisch karakter hij kennelijk toeschrijft aan de causale vermogens van het menselijk brein, evenals aan het fenomeen intentionaliteit. Welke causale vermogens kan het brein hebben, behalve de causale vermogens van individuele neuronen, wiens gedrag nagebootst kan worden? Eigenlijk komt het er nu volgens hem op neer dat 'echte' intelligentie, met werkelijk begrip, slechts weggelegd is voor mensen. Wat er dan precies zo speciaal is aan het menselijk brein, en wat intentionaliteit is (behalve dat mensen het hebben en thermostaten bijvoorbeeld niet), dat kan Searle niet zeggen, en er mag ook niet aan getwijfeld worden (zoals Dennett doet door te stellen dat het zou kunnen dat de meerderheid van de mensheid geen intentionaliteit meer heeft omdat deze geen voordeel geeft voor natuurlijke selectie).

Het gedachtenexperiment rust op de intuïtie van de lezer over de intentionaliteit van mensen, en het gevoel dat er *iets* mist in andere gevallen zoals bij digitale computers. Een emergentist of epiphenomenalist zal hier niet van onder de indruk zijn.

1.5 Bronnen

Searle, John (1980), "Minds, Brains and Programs", Behavioral and Brain Sciences 3(3): 417-457
`http://www.bbsonline.org/documents/a/00/00/04/84/`