

Dit is een pre-print versie van het volgende artikel:

Uithol, S. & Brey, P. (2006). 'Hubert Dreyfus,' *Kritisch Denkerslexicon* 36, pp. 1-23. Ed. H. Achterhuis and J. Sperna Weiland. Veen.

Hubert Dreyfus

Biografie

Hubert Lederer Dreyfus is in 1929 in Terre Haute, Indiana geboren. In 1964 promoveert Dreyfus in de filosofie aan de universiteit van Harvard, waarna hij filosofie onderwijst aan het 'Massachusetts Institute of Technology'. Hier maakte hij kennis met de ontwikkelingen van de Artificiële Intelligentie (AI). Hierdoor geprikkeld zoekt hij contact met zijn broer Stuart Dreyfus, die als computerspecialist werkzaam is bij de RAND-organisatie, een breed georiënteerd onderzoeksinstituut. Hubert Dreyfus wordt hier als filosofisch consulent aangesteld, met als taak hun nieuw AI-onderzoek onder leiding van Allen Newell en Herbert Simon te evalueren. In 1965 schrijft hij het vernietigende rapport *Alchemy and Artificial Intelligence*, een felle kritiek op het werk van de onderzoekers Newell en Simon. Dit is de aanzet van een reeks kritieken op de Artificiële Intelligentie, met *What Computers Can't Do: A Critique of Artificial Reason* uit 1972 als het belangrijkste en bekendste werk. Hiervan is in 1992 een nieuwe editie onder de naam *What Computers Still Can't Do* verschenen.

Wat Dreyfus' werk over het AI-onderzoek bijzonder maakt is dat hij denkers uit de geschiedenis van de filosofie effectief weet in te zetten als critici van de AI. Het gaat hierbij om Europese fenomenologen als Martin Heidegger en Maurice Merleau-Ponty, die meer gericht zijn op de aard van de mens, en minder op wetenschap en technologie.

Naast zijn werk over Artificiële intelligentie is Dreyfus belangrijk in het verbreiden van de ideeën van Heidegger en Michel Foucault in de Verenigde Staten. Hij heeft naast zijn kritieken op de AI belangrijke verhandelingen over Heidegger, Merleau-Ponty, Husserl, Foucault en Kierkegaard afgeleverd. Zijn boek *Being in the World* uit 1991 wordt gewaardeerd als een van de beste en helderste uiteenzettingen van het vroege werk van Heidegger. Recent heeft Dreyfus ook enkele invloedrijke kritieken gepubliceerd over het Internet. Dit artikel richt zich vooral op Dreyfus' kritieken van de AI en van het Internet.

Kritische beschouwing

Symbolische AI

Dreyfus heeft vooral naam gemaakt met zijn kritiek op de symbolische artificiële intelligentie. Deze vorm van AI-onderzoek wordt zo genoemd vanwege het centrale uitgangspunt ervan, namelijk dat intelligentie een kwestie is van het verwerken en manipuleren van symbolen volgens vaststaande, formele regels. Dit uitgangspunt is gebaseerd op drie aannamen. Ten eerste dat intelligente processen, zoals waarnemen, redeneren, rekenen en taal gebruiken vormen van informatieverwerking zijn. Dit is het opnemen van informatie uit de omgeving, het verwerken of manipuleren van deze informatie, en het geven van een respons.

Gerepresenteerde informatie

Als intelligentie op informatieverwerking berust, dan dient de vraag zich aan wat bij intelligente systemen het karakter van deze informatie is, en op welke manier deze informatie ‘verwerkt’ wordt. Binnen de symbolische AI wordt aangenomen dat informatie gerepresenteerd moet worden om verwerkt te kunnen worden. Om informatie in zich op te nemen moet het systeem beschikken over een medium waarop deze informatie kan worden ingeschreven. Informatieverwerkende systemen, zo wordt verondersteld, maken gebruik van interne representaties, die een aan het systeem aangepaste vorm aannemen. Zo wordt in het menselijk denken gebruik gemaakt van een systeem van interne, mentale representaties, waarop gedachten, herinneringen en waarnemingen zijn ingeschreven.

Symbolische representatie

De volgende aanname over de aard van deze representaties is de meest kenmerkende voor de klassieke AI: zij stelt dat de interne representaties van intelligente systemen *symbolisch* van aard zijn. Men had bijvoorbeeld ook kunnen aannemen dat interne representaties *iconisch* zijn: dat ze informatie dragen doordat ze een uiterlijke gelijkens bezitten met datgene waarnaar ze verwijzen, zoals een schilderij informatie verschafft over het geportretteerde door erop te lijken. Maar de klassieke AI gaat ervan uit dat de representaties meer overeenkomen met natuurlijke taal. Taal is symbolisch: taaltekens hebben geen enkele gelijkens of inherente verwijzing naar dat wat ze representeren; ze zijn volstrekt arbitrair. Zo verwijst het woord ‘hond’ naar honden, maar lijkt het zelf niet op een hond.

Symbolische representaties hebben een belangrijk voordeel boven iconische representaties. Zo zijn iconische representaties niet in staat om abstracte zaken als bijvoorbeeld logische operatoren (‘en’, ‘of’) of temporele noties (‘morgen’, ‘volgende week’) te representeren, omdat deze geen waarneembare structuur hebben. Ook zijn symbolische representaties veel makkelijker combineerbaar dan iconische, en kan men met een eindig aantal symbolen een bijna oneindig aantal verschillende informatie-inhoudende representeren. Dit is te zien in de natuurlijke taal, waar een eindig aantal woorden kan worden gecombineerd tot een in praktisch oneindig aantal verschillende zinnen. Bovendien lijken symbolen een prominente rol te spelen in het soort cognitieve taken die doorgaans als de meest intelligente worden gezien, zoals het maken van wiskundige berekeningen en logisch redeneren.

Formele verwerking

De symbolen worden in de symbolische AI uitsluitend verwerkt op basis van hun formele eigenschappen, dat wil zeggen, die eigenschappen die ze tot symbool maken. De betekenis

van een symbool speelt dus geen directe rol in de verwerking ervan. Zo zal een computer het symbool 'hond' op een bepaalde manier verwerken omdat hij de vorm van het symbool herkent, en niet omdat hij inzicht heeft in de betekenis van dit symbool. Een informatieverwerkend systeem beschikt over regels waarmee deze symbolen geïnterpreteerd en verwerkt worden. Dit zijn formele regels, omdat ze slechts aangrijpen op de formele eigenschappen van deze symbolen. Deze aanname is gebaseerd op de intuïtie dat het toepassen van regels een rol lijkt te spelen bij veel intelligente taken. Zo lijkt het interpreteren van taal kennis van de grammaticaregels in te houden en lijkt logisch redeneren het toepassen van regels van de logica in te houden. Merk op dat men er hierbij vanuit gaat dat de kennis die voor intelligentie vereist is theoretisch is. Iets kennen en begrijpen betekent dat men een abstracte, in regels gevatte, symbolische theorie bezit waarin men het verschijnsel kan vangen.

De door AI-onderzoekers geschetste theorie van intelligentie kan men formuleren en beredeneren zonder enige verwijzing naar de aard en mogelijkheden van de digitale computer. Het zal echter duidelijk zijn dat de ontwikkeling van de computer deze theorie van intelligentie aanmerkelijk aantrekkelijker maakt. Een digitale computer is immers op te vatten als een informatieverwerkend systeem, dat bovendien gebruik maakt van symbolische representaties (combinaties van nullen en enen), en deze verwerkt volgens formele regels (het eveneens symbolisch gerepresenteerde programma van de computer). Het bestaan van de computer biedt de kans om aannamen over het symbolische en regelgestuurde karakter van intelligentie wetenschappelijk te onderzoeken, en doorzichtige modellen van intelligente processen te construeren. De bovengenoemde aannamen over intelligentie bieden dus uitzicht op een potentieel vruchtbaar wetenschappelijk onderzoeksprogramma dat lijkt te kunnen leiden tot interessante technische toepassingen. De aanvankelijke successen van de symbolische AI in het ontwerpen van intelligente computerprogramma's bieden hierbij nog een extra legitimatie.

Rationalistische traditie

Dreyfus benadrukt dat de opvattingen over de symbolische aard van intelligentie niet nieuw zijn, maar slechts nieuwe reïncarnaties zijn van oude ideeën die met name gevonden kunnen worden in wat wel de *rationalistische* traditie in de filosofie wordt genoemd. Plato legde de basis voor deze theorie door te stellen dat wijsheid bestaat uit het kunnen formuleren van kennis in expliciete definities. Hij meende dat het mogelijk was een systeem van theoretische, objectieve principes te vinden die, zoals de basisprincipes in de geometrie, gebruikt konden worden om op rationele basis de werkelijkheid te verklaren en handelingen te rechtvaardigen. Dergelijke opvattingen over intelligentie vindt men terug bij veel andere rationalistisch ingestelde denkers zoals Descartes, Leibniz, Kant, en Husserl, maar in mindere mate ook bij empiristen als Locke en Hume, en meer recent bij wetenschappers zoals de beroemde linguïst Noam Chomsky, de psycholoog en filosoof Jerry Fodor, en de verschillende representanten van de symbolische AI.

Psychologische vooronderstelling

Volgens Dreyfus zijn er met name drie voor het rationalisme karakteristieke opvattingen in de symbolische AI terug te vinden. De eerste noemt hij de *psychologische vooronderstelling*, de eerder genoemde aanname van de symbolische AI dat menselijke intelligentie een kwestie is

van het manipuleren van symbolen volgens formele regels. Deze aanname maakt het in theorie mogelijk een computer zo te programmeren dat hij denkt als een mens. Niet al het onderzoek in de symbolische AI maakt echter deze psychologische veronderstelling; het is vooral een expliciet uitgangspunt van de Cognitieve Simulatie-benadering: AI-onderzoek dat als expliciet doel het modelleren van menselijke cognitieve processen heeft. In alle variëteiten van symbolische AI wordt echter wel een tweede aanname gedaan: de epistemologische vooronderstelling.

epistemologische veronderstelling.

Deze aanname stelt dat alle kennis formaliseerbaar is, dat wil zeggen, dat alles wat door mensen begrepen kan worden, uitgedrukt kan worden in context-onafhankelijke, formele regels of definities. De waarheid van deze veronderstelling zou garanderen dat het project van de AI zelfs als de psychologische veronderstelling onjuist is toch kan slagen, omdat een geformaliseerde versie van informele menselijke kennis en inzicht dezelfde informatieve waarde heeft als het ongeformaliseerde origineel. Een computer is dan misschien niet in staat om menselijke denkprocessen te simuleren, maar wel om menselijk intelligent gedrag te reproduceren.

ontologische veronderstelling

Zowel de epistemologische als de psychologische veronderstelling zijn tenslotte mede gebaseerd op de ontologische veronderstelling: de door de mens kenbare werkelijkheid heeft een formaliseerbare structuur, doordat deze is opgebouwd uit een serie objectieve, onafhankelijk van elkaar kenbare feiten. Als de werkelijkheid deze structuur niet had, dan zou het onwaarschijnlijk zijn dat zij kenbaar is met behulp van een verzameling context-onafhankelijke, formele regels, zoals in zowel de epistemologische als psychologische veronderstelling wordt verondersteld.

Op de rationalistische traditie zijn echter belangrijke filosofische kritieken verschenen, die volgens Dreyfus dus ook op de symbolische AI van toepassing zijn. Dreyfus is met name onder de indruk van de meer recente anti-rationalistische perspectieven van Heidegger, Merleau-Ponty en Wittgenstein. Hij maakt gebruik van deze perspectieven in zijn kritiek op de symbolische AI en haar drie centrale veronderstellingen, en in zijn ontwikkeling van een alternatieve theorie van intelligentie.

Dreyfus' belangrijkste kritiek is gericht op de epistemologische veronderstelling dat intelligent gedrag reproduceerbaar is door menselijke kennis te formaliseren (in regels uit te drukken) en in een computer op te slaan. Deze vooronderstelling is terug te vinden in alle vormen van symbolische AI. Dreyfus' argument tegen deze veronderstelling is dat ofschoon het wel mogelijk is om (menselijke) kennis bij benadering te beschrijven, het niet mogelijk lijkt om op grond van deze beschrijvingen deze kennis te reproduceren. Het hebben van kennis vooronderstelt volgens Dreyfus de vaardigheid om deze kennis in relevante situaties te kunnen reproduceren of toepassen, in redeneringen, handelingen en communicatie. Het weten dat vuur heet is houdt bijvoorbeeld in dat deze kennis op relevante momenten wordt toegepast in het nadenken over vuur of het omgaan met vuur.

Het toepassen van geformaliseerde, in regels weergegeven kennis lijkt op een belangrijk probleem te stuiten. Wil een computer die is uitgerust met een verzameling formele kennisregels deze regels kunnen toepassen op een nieuw gegeven, bijvoorbeeld een ingetypte

zin, een rekensom of een via een camera waargenomen beeld, dan moet eerst bepaald worden of deze regels wel op het nieuwe gegeven van toepassing zijn. Formele regels zijn er echter op gericht zo min mogelijk acht te slaan op context en slechts oog te hebben voor de gegevens die in de regel als aangrijpingspunt worden geformuleerd. Om regels toch gevoelig te maken voor context moeten allerlei contexten worden geformuleerd, of aparte toepassingsregels worden geformuleerd, maar dit proces lijkt zonder einde te zijn.

Stel bijvoorbeeld dat geprobeerd wordt een computer te programmeren om taal te begrijpen. De computer zou dan de betekenis van stukken tekst kunnen achterhalen door de complexe betekenis van deze tekst op te bouwen met behulp van de interpretatieregels en de grammaticaregels. Echter, sommige woorden zijn dubbelzinnig. Als de computer het woord 'heet' krijgt aangeboden, zal hij meestal de interpretatieregel 'Als iets heet is, heeft het een hoge temperatuur' kunnen toepassen, maar als het gaat om hete gerechten kan de interpretatieregel 'Als iets heet is, is het gepeperd en brandt het in de mond' gelden. Om te weten welke van deze twee interpretatieregels moet worden toegepast, zijn andere gegevens relevant, bijvoorbeeld of er in de tekst wordt verwezen naar voedsel. Er moeten dus regels worden opgesteld voor de juiste *toepassing* van de interpretatieregels. Ook op deze toepassingsregels zullen echter weer uitzonderingen bestaan. Zo kan een tekst gaan over hete Thaise gerechten, waarbij echter duidelijk uit de context blijkt dat op de hoge temperatuur ervan wordt gedoeld.

Gezond verstand

Mensen kunnen wél moeiteeloos gegevens interpreteren vanuit hun context. Mensen, concludeert Dreyfus, hebben 'gezond verstand', waardoor sommige interpretaties zinnig lijken en andere niet. Computers hebben geen gezond verstand en komen daardoor vaak tot onzinnige interpretaties. Het bijbrengen van gezond verstand aan computers lijkt Dreyfus de grootste uitdaging voor de symbolische AI. Hij noemt dit het gezond-verstandprobleem van de symbolische AI, en is van mening dat dit probleem om bovengenoemde redenen onoplosbaar is. Dreyfus werkt het begrip gezond verstand helaas niet verder uit, en laat zijn stelling daarmee leunen op dit 'common sense' begrip, zonder duidelijk te maken hoe de mens wel in staat is gezond verstand te hebben, en waarom symbolische AI hier van verstoken zal blijven.

Er zijn domeinen waar computers goed presteren: formele en abstracte domeinen waar de identiteit van de elementen direct af te lezen is, waardoor een gebrek aan gezond verstand geen probleem is. Voorbeelden hiervan zijn schaak, wiskunde en formele logica. Echter, alledaagse probleemsituaties waarin intelligent gedrag vereist is, lijken echter een heel andere structuur te hebben dan deze kunstmatige 'werelden'. In alledaagse situaties is het doel vaak niet eens duidelijk. Bovendien laat het probleem zich niet gemakkelijk formaliseren, zelfs als het doel helder zou zijn, omdat niet bij voorbaat duidelijk is welke feiten potentieel relevant zijn voor het vinden van een oplossing. Samenvattend lijken er goede argumenten te bestaan tegen de epistemologische veronderstelling dat intelligent gedrag reproduceerbaar is door middel van een systeem van formele regels en symbolen. Bovendien lijkt menselijke intelligentie niet op deze manier te werken.

Alternatieve theorie

In zijn alternatieve theorie van intelligentie verdedigt Dreyfus het uitgangspunt dat mensen bij intelligent gedrag meestal geen regels toepassen en meestal zelfs geen gebruik maken van interne representaties.

Gesitueerde intelligentie

Intelligentie is volgens Dreyfus gesitueerd; zij wordt mede bepaald door de situatie waarin men zich bevindt. Het inzicht waarop intelligent gedrag is gebaseerd wordt lokaal vanuit een concrete situatie opgebouwd met behulp van informatie die door deze situatie direct verschaft wordt, zonder dat hiervoor regels of interne representaties nodig zijn. Deze visie is vooral gestoeld op de filosofie van Heidegger en in mindere mate op die van Wittgenstein en Merleau-Ponty.

Volgens Dreyfus wordt de wereld door de natuurwetenschappen geïnterpreteerd als een materiële, van de mens onafhankelijke structuur die inherent betekenisloos, ruimtelijk gescheiden van de mens is, zodat geen directe, ongemedieerde ervaring ervan mogelijk is. Dreyfus ontkent niet de waarde van het dit perspectief, maar benadrukt dat er nog een ander perspectief mogelijk is. Dit is een fenomenologisch perspectief dat de ervaring van de mens als uitgangspunt neemt bij het beschrijven van de wereld. Deze 'menselijke wereld' is een wereld die niet geheel objectief is, maar gevuld is met ervaren structuren, zoals geuren, gevoelens, frustraties, bedreigingen, hindernissen en doelen. Deze wereld is echter ook niet geheel subjectief, in die zin dat het soort patronen die wij in deze wereld leren ontwaren niet geheel willekeurige constructies van onze geest zijn; geuren en hindernissen zijn geen zaken die we zomaar bedenken, maar die zich in onze interactie met de wereld in onze ervaring manifesteren. Het is nu in deze noch geheel objectieve, noch geheel subjectieve wereld, waar de mens actief is en leert waarnemen, handelen en denken. Zo is de wereld van een pasgeboren baby nog grotendeels ongestructureerd, maar bevinden zich in de wereld van een volwassen mens talloze, in de loop van jaren uitgekristalliseerde structuren. De wereld evolueert echter mee met deze activiteiten, omdat mede in en door deze activiteiten zich steeds weer nieuwe structuren manifesteren.

Dreyfus betoogt dat bij mensen de ervaring van de wereld-in-zijn-geheel altijd voorafgaat aan een ervaring van afzonderlijk te onderscheiden elementen. De specifieke elementen worden onderscheiden en ervaren vanuit een meer algemene ervaring van betekenis en zin. Hierdoor staan deze elementen altijd in een zinvol betekenisverband met hun context. Intelligent gedrag vereist dat een mens in de situatie waarin hij of zij zich bevindt een betekenisvolle structuur ontwaart die handelingen voorschrijft die gegeven de situatie zinvol zijn. Deze structuur is het lokale product van de behoeften, handelingen, en resulterende waarnemingen van deze mens. De verschillende elementen in deze structuur zijn voor hun betekenis afhankelijk van deze totale structuur. Vanuit deze directe betekenisvolheid volgt een handelingspatroon als automatisch. Dit handelingspatroon kan vrijwel automatisch vanuit de waargenomen situatie worden gegenereerd, omdat deze situatie reeds gestructureerd is op een manier die hem 'hanteerbaar' maakt, dat wil zeggen: in termen van zinvolle handelingen.

De betekenisvolle structuur die in een situatie wordt waargenomen is dus niet een structuur die volgens een aantal vaste regels wordt opgebouwd uit afzonderlijke, context-onafhankelijke elementen, zoals de symbolische AI veronderstelt. Het is precies omgekeerd: de globale, holistische structuur die aan een situatie wordt toegekend maakt het mogelijk om

vervolgens elementen uit deze structuur door een proces van abstractie te representeren als afzonderlijke objecten en feiten, die onder regels te vangen zijn. Over hoe precies het toekennen van een structuur aan een situatie in zijn werk gaat blijft Dreyfus echter onduidelijk.

Belichaamde intelligentie

Dreyfus' opvatting over de gesitueerdheid van menselijk handelen vormt de ene helft van zijn theorie van menselijke intelligentie. De andere helft wordt gevormd door zijn opvatting dat intelligentie belichaamd is, dat wil zeggen, een lichaam vereist (Dreyfus 1967, 1972, 1996). Deze uiteenzetting, die vooral geïnspireerd is op de filosofie van Merleau-Ponty, staat niet geheel los van de eerste en is, net als zijn uiteenzetting van zijn ideeën over de gesitueerdheid van intelligentie, onduidelijk en schetsmatig. Het is met name niet duidelijk of Dreyfus bedoelt dat intelligentie iets is dat noodzakelijkerwijs over een heel lichaam gedistribueerd is en dus niet alleen in de hersenen of de geest is gelokaliseerd, of dat intelligentie wel kan bestaan zonder een lichaam, maar alleen ontwikkeld kan worden met behulp van een lichaam. Voor beide mogelijkheden lijken neuropsychologische aanwijzingen te bestaan. Dreyfus lijkt tegenwoordig vooral de laatst mogelijkheid te huldigen. Een zich ontwikkelende abstracte intelligentie zou direct kunnen voortbouwen op sensomotorische vaardigheden als patroonherkennen, het visueel scannen van beelden, het mentaal groeperen en manueel manipuleren van voorwerpen, door ze toe te passen in meer abstracte domeinen. Dreyfus verwijst in zijn recente werk naar studies van Mark Johnson die heeft proberen aan te tonen dat abstracte begrippen en abstracte logica uiteindelijk te herleiden zijn tot concrete, sensomotorische structuren.

Als intelligentie inderdaad gesitueerd en belichaamd is, dan lijkt het niet mogelijk voor digitale computers om het brede scala van menselijke intelligentie te bezitten, omdat ze niet belichaamd zijn en niet beschikken over een menselijke wereld zoals die hierboven geschetst is. De intelligentie van computers lijkt beperkt te blijven tot het verrichten van taken in voorgedefiniëerde formele domeinen en zal falen in een complexe, menselijke wereld.

Neurale netwerken, connectionisme

Sinds het begin van de tachtiger jaren is er, mede door de gebleken tekortkomingen van het symbolische paradigma, een rivaliserend paradigma ontstaan in de AI, dat bekend staat onder de naam neurale netwerken of connectionisme. Neurale netwerken-AI wordt door de meeste wetenschappers gezien als een radicaal alternatief voor symbolische AI. De neurale-netwerkenbenadering neemt afstand van het idee dat intelligent gedrag voortkomt uit het manipuleren van symbolen volgens formele regels en laat zich inspireren door de structuur en werking van de menselijke hersenen. Deze zijn opgebouwd uit neuronen (zenuwcellen): kleine informatieverwerkende systeemjes. Neuronen ontvangen prikkels, van andere neuronen of soms direct van zintuigcellen waarmee ze in contact staan en reageren hierop door zelf elektrochemische prikkels af te geven aan andere zenuwcellen, of soms ook spieren en klieren, waar ze op aangrijpen via zenuwuitlopers. Of een neuron zelf prikkels afgeeft wordt bepaald door een 'programma' in de zenuwcel volgens welke hij impulsen die hij zelf te verwerken krijgt bij elkaar optelt (of soms aftrekt). Boven een bepaalde drempelwaarde leidt deze optelsom tot een reactie: de cel geeft zelf een impuls af aan zijn omgeving.

Neuronen zijn dus te begrijpen als verwerkingseenheden (processors) met tamelijk eenvoudige input/output functies.

Intelligentie is nu voornamelijk een product van de connecties die neuronenaangaan (vandaar de naam connectionisme). Neuronen ontwikkelen zich door connecties met de omgeving aan te gaan of te verbreken, of te versterken of te verzwakken, afhankelijk van de mate waarin ze zelf geprikkeld worden. Bij iemands geboorte zijn de connecties die zenuwcellenaangaan nog tamelijk willekeurig, maar door de wisselwerking met zijn omgeving wijzigen de connecties van zijn zenuwcellen zich zo dat het door het zenuwstelsel geïnstigeerde gedrag steeds intelligenter en succesvoller wordt. Een zenuwstelsel leert dus doordat de connecties tussen zenuwcellen door ervaring gemodificeerd worden.

Artificiële neurale netwerken blijken verbazingwekkend goed in staat om bepaalde intelligente taken te verrichten, zoals het herkennen van patronen, het categoriseren van gegevens en het coördineren van handelingen. Zo zijn er bijvoorbeeld netwerken die vanuit verschillende invalshoeken gezichten kunnen herkennen en die geschreven teksten goed leren uitspreken. Neurale netwerken zijn echter vooral goed in het uitvoeren van taken die 'lagere' vormen van intelligentie vereisen, zoals patroonherkenning, categorisatie en motorische sturing. Het is tot nu toe moeilijker gebleken om neurale netwerken wiskundige of logische problemen te laten oplossen, wat juist het soort taken is waarin de symbolische AI de meeste successen heeft geboekt.

Hubert Dreyfus stelt dat de uitgangspunten van neurale netwerken-AI goed verenigbaar zijn met zijn eigen visie op intelligentie (Dreyfus 1992). De neurale-netwerkenbenadering stapt af van het rationalistische idee dat intelligentie een kwestie is van het manipuleren van symbolen en het toepassen van regels. Kennis is in neurale netwerken niet een kwestie van het bezitten van expliciete representaties, maar van het hebben van de juiste verbindingen tussen zintuigen en spieren. Het hebben van kennis is in principe het hebben van een vaardigheid: het is meer weten hoe je iets moet doen dan weten dat een bewering waar is. Intelligente processen zijn vaak holistisch en intuïtief. Bovendien is neurale netwerken-AI goed verenigbaar met het uitgangspunt dat intelligentie een lichaam vereist en gesitueerd is; hogere taken worden vaak opgebouwd uit lagere en intelligentie wordt gezien als iets dat zich ontwikkelt door interactie met een omgeving. Neurale netwerken-AI lijkt daarom volgens Dreyfus' eigen criteria alles mee te hebben wat nodig is om echte artificiële intelligentie te vervaardigen.

Toch is Dreyfus uiteindelijk pessimistisch over de mogelijkheid van neurale netwerken om menselijke intelligentie te benaderen. Dit ligt niet aan de uitgangspunten van de neurale netwerken-AI maar aan de grote complexiteit van menselijke intelligentie. De menselijke hersenen tellen circa honderd miljard neuronenaan. Wil menselijke intelligentie benaderd worden, dan zal ook dit aantal benaderd moeten worden en zouden neurale netwerken moeten bestaan uit miljarden nodes, in plaats van enkele tientallen of honderdtallen.

Afgezien van het feit dat dit momenteel technisch moeilijk haalbaar lijkt, is er nog de vraag hoe een dergelijk netwerk alle relevante kennis waarover een normaal mens beschikt aangeleerd krijgt. Het verwerven van deze kennis lijkt te vereisen dat zo'n netwerk net zo'n leertraject doormaakt als een volwassen mens achter zich heeft. Maar dit vereist, zoals is betoogd in de vorige sectie, dat het neurale netwerk belichaamd is. Het lijkt er dus op dat neurale netwerken die even intelligent kunnen generaliseren als mensen alleen verkregen

kunnen worden door neurale netwerken te bouwen met de complexiteit van de menselijke hersenen, en deze netwerken in te bouwen in kunstmatige lichamen die een ontwikkelingstraject doorgaan zoals dat van opgroeiende mensen. Het creëren van dergelijke androïde levensvormen is tot nu toe sciencefiction.

AI als technologie

Hoewel de eventuele praktische toepasbaarheid van AI-onderzoek al vroeg in het achterhoofd speelde van AI-onderzoekers en hun financiers, profileerde de AI zich aanvankelijk vooral als wetenschap en zijn er tot aan het eind van de jaren zeventig nauwelijks interessante toepassingen van AI-onderzoek geweest. Sinds het midden van de jaren tachtig is de AI echter steeds meer het karakter gaan aannemen van een technologie. Inmiddels is AI als technologie een miljardenindustrie geworden en heeft sinds de late zeventiger jaren een stroom van producten opgeleverd, zoals schaakcomputers en expertsystemen. Vooral sinds de negentiger jaren is er een opkomst te zien van conventionele producten die met artificiële intelligentie worden uitgerust, zoals 'intelligente' stofzuigers, wasmachines en videocamera's 'intelligente' regelsystemen in de industrie, 'intelligente' computersoftware, zoals besturingsystemen die hun eigen gedrag afstemmen op het gebruikspatroon van de gebruiker en 'intelligente' zoekprogramma's voor elektronische kennisbestanden. De grens tussen AI-onderzoek en ander technologisch onderzoek (met name in informatica en elektrotechniek) is door de gerichtheid op dit soort toepassingen aan het vervagen.

De massale toepassing van intelligente computersystemen in de maatschappij brengt nieuwe filosofische vragen met zich mee, met name naar de ethische implicaties ervan. De belangrijkste ethische vraagstukken met betrekking tot intelligente computersystemen liggen bij expertsystemen. Expertsystemen, waarvan de eerste in het midden van de jaren zeventig ontwikkeld werden, zijn computersystemen die bedoeld zijn om taken over te nemen van experts in het specialistische domein waarin deze actief zijn. Toepassingen van expertsystemen liggen onder andere in de medische wereld, het recht, de industrie, de wis- en natuurwetenschappen, financiële planning en accountancy. Zo zijn er expertsystemen vervaardigd om ziektebeelden te diagnosticeren en behandelingsmethoden aan te bevelen, fouten op te sporen in vliegtuigmotoren, gebieden te identificeren waar mogelijk mineralen aanwezig zijn voor mijnbouw, portfolio's samen te stellen voor investeerders, vast te stellen of personen recht hebben op een werkloosheidsuitkering en een strafmaat te bepalen voor veroordeelde wetsovertreders.

Expertsystemen zijn meestal gebouwd volgens de uitgangspunten van de symbolische AI. Men probeert deze systemen de benodigde specialistische kennis te geven door experts te interviewen en te trachten hun vaak ongeverbaliseerde en intuïtieve kennis expliciet te maken. Dit leidt tot een lijst van (vaak duizenden of tienduizenden) feiten en heuristieken (regels volgens welke experts worden geacht te redeneren), die vervolgens in een computerprogramma worden vertaald. Daarna worden de prestaties van het systeem vergeleken met de prestaties van een menselijke expert. Als het systeem voldoende lijkt te presteren, kan het in gebruik genomen worden.

Aanvankelijk is Dreyfus ondanks zijn kritiek op de symbolische AI nog tamelijk optimistisch over de mogelijkheden van expertsystemen. Dreyfus heeft altijd gesteld dat computers goed

kunnen presteren in geformaliseerde domeinen die weinig ‘gezond verstand’ vereisen. Het soort kennis dat door experts is verworven, zoals schaakgrootmeesters lijkt vaak terug te voeren op aangeleerde formele regels, zoals de regels van het schaakspel en lijkt weinig ‘gezond verstand’ of alledaagse kennis te vereisen. Dreyfus meende dat computers in dit specialistische kennisdomeinen dan ook goede successen zouden kunnen boeken. Later is hij hier op teruggekomen. Belangrijkste aanleiding daarvoor vormde een studie, door hem uitgevoerd samen met zijn broer Stuart, van de manier waarop mensen expertise ontwikkelen in een bepaald gebied (Dreyfus & Dreyfus 1986). Deze studie lijkt aan te tonen dat mensen in vroege leerstadia gebruik maken van regels, maar bij het bereiken van expertise deze regels inmiddels vervangen hebben door een intuïtieve en holistische manier van probleemoplossen. Een schaakgrootmeester, bijvoorbeeld, past bij het schaken geen regels meer toe, zoals beginners dat doen, maar ‘ziet’ in een enkele blik een bordpositie en een aantal mogelijke tegenzetten. Zijn expertise rust niet in opgeslagen feiten en regels, maar in zijn herinneringen van situaties in het verleden die hij succesvol tegemoet is getreden. Eerder aangeleerde regels (zoals ‘Zorg dat je de koningin vroeg in het spel kunt gebruiken’ en ‘Een toren is meestal meer waard dan een loper’) zijn vervangen door een kennisbestand van tienduizenden globaal waargenomen bordpatronen en bijbehorende tegenzetten.

Aangeleerde regels bieden vooral een hulpmiddel voor de beginneling en de halfgevoerde in een bepaald kennisdomein, een beginstructuur die een versimpelde kijk op de werkelijkheid geeft, maar een handvat biedt van waaruit de eigenschappen en vereisten van specifieke gevallen aangeleerd kunnen worden. Omdat volgens Dreyfus (*contra* de ontologische veronderstelling) de realiteit geen formele structuur heeft die in regels kan worden gevangen, bestaat expertise uiteindelijk in het kennen van en kunnen omgaan met talloze aparte gevallen. Expertsystemen die wel uitgaan van de formaliseerbaarheid van de kennis van experts zullen dus volgens Dreyfus het niveau van echte expertise nooit halen. Het is dan ook niet te rechtvaardigen om ze in te zetten voor taken die expertise vereisen. Dreyfus is echter wel overtuigd dat expertsystemen nuttig kunnen zijn op het niveau van *competentie*: een prestatieniveau dat een beginnersniveau voorbijstreeft en vergelijkbaar is met dat van een gevorderde student. Een vraag die Dreyfus hier echter niet aansnijdt, is hoe beslist kan worden of een bepaalde taak expertise of slechts competentie vereist.

Intelligent Tutoring Systems

Een tweede type intelligente computersystemen dat door Dreyfus wordt besproken, dat nauw verwant is aan expertsystemen, bestaat uit intelligente onderwijssystemen (‘intelligent tutoring systems’ of ITS-en), die worden ingezet in computerondersteund onderwijs. Intelligente onderwijssystemen zijn computerprogramma’s die taken overnemen van leraren door leerlingen individueel te onderrichten. Zij zijn meestal niet als totale vervanging van de leraar bedoeld, maar als aanvulling op het lesgeven. Er is een belangrijk verschil tussen het gebruik van een computer als ITS en het gebruik ervan in andere functies, zoals tekstverwerker, tekenbord of elektronische encyclopedie. Voor zulke toepassingen wordt gebruik gemaakt van ‘onintelligente’ computerprogramma’s. Een ITS is echter een programma dat intelligentie pretendeert, omdat het pretendeert sommige van de vaardigheden te bezitten van een professionele leraar.

Intelligente onderwijssystemen kunnen leerlingen op twee manieren helpen. Een eerste, eenvoudig, type onderwijssysteem biedt vraagstukken of problemen aan waarna de leerling het goede antwoord moet geven. Het kan dan bijvoorbeeld gaan om spellingsoefeningen of oefeningen in algebra. De computer heeft dan een vaardigheid in het genereren van nieuwe vragen of problemen en in het evalueren van de antwoorden van de leerling. Dreyfus ziet weinig bezwaren tegen dit type ITS. Hij meent dat computers bij uitstek geschikt zijn om leerlingen met behulp van voorbeelden en oefeningen kennis en vaardigheden bij te brengen in een bepaald domein. Het enige gevaar met deze toepassingen is dat ze, omdat ze zo goed werken, te veel benadrukt kunnen gaan worden in het leerproces, ten koste van andere leervormen.

Een meer verregaand type ITS neemt een actief begeleidende rol aan door adviezen en aanwijzingen te geven, uit te leggen wat de student verkeerd doet en de aangeboden vraagstukken en het tempo af te stemmen op de individuele leerling. Dit type ITS wordt gebruikt bij het aanleren van meer complexe kennis of vaardigheden, waarbij de taak bijvoorbeeld is om bepaalde theorieën en begrippen te leren beheersen en ze te kunnen toepassen in concrete situaties.

Dreyfus' eerste bezwaar tegen zulke ITS-en is dat ze ongeschikt zijn om leerlingen te helpen expertise in een bepaald domein te ontwikkelen. Om dit te kunnen moet een computersysteem eerst zelf expertise bezitten. Maar zoals al eerder betoogd is, is het niet mogelijk om computersystemen die volgens de symbolische AI zijn geprogrammeerd expertise mee te geven. Daarom zijn ITS-en hooguit geschikt om een beperkte mate van competentie in een gebied aan te leren. Zij zijn vooral geschikt in vroege leerstadia, omdat daarin nog regels moeten worden aangeleerd. Het zou echter desastreus zijn om ITS-en ook in latere leerstadia te gebruiken, omdat ze gebruik maken van regels, en dus in het doceren deze regels steeds aan de leerling zullen opleggen. Omdat expertise juist wordt verworven doordat regels op een bepaald moment worden losgelaten, zal zo'n ITS mensen alleen maar verhinderen om echte expertise te verwerven.

Wanneer gekozen wordt om ITS-en slechts bij beginners en halfgevoerden in te zetten is er nog een ander groot probleem. Dit bestaat eruit dat ITS-en om een goede docent te zijn moeten beschikken over grote didactische vaardigheden. Een goede docent is niet alleen iemand met vakkennis, het is ook iemand die aansluiting weet te vinden bij de kennis en vaardigheden waarover de leerling reeds beschikt en die zijn of haar manier van onderwijzen daaraan weet aan te passen. Een natuurkundedocent moet bijvoorbeeld een inzicht hebben in de naïeve opvattingen over de werking van de natuur die studenten met zich meebrengen en hierop kunnen inspringen. Zo zal ook een ITS een dergelijk didactisch inzicht en aanpassingsvermogen moeten hebben.

Het probleem is echter dat in een ITS de veronderstelde kennis en vaardigheden van de leerling alleen in termen van een aantal symbolen en regels kunnen worden uitgedrukt. De ITS neemt hiermee impliciet aan dat de leerling die hij onderwijs geeft een regelvolgend, symbool-manipulerend, rationeel wezen is. In feite is de leerling echter een belichaamd wezen dat een menselijke wereld bewoont waarin de ITS zich zou moeten kunnen verplaatsen om werkelijk te begrijpen vanuit welke beginsituatie de leerling probeert te leren. Omdat ze dit niet kunnen zijn ITS-en ongeschikt om leerlingen te helpen met het zien van de onderlinge

samenhangen die nodig zijn voor het leren beheersen van een nieuw kennisdomein. Samenvattend is het probleem met ITS-en dat ze bij onderwijs aan gevorderden vaak op vakinhoudelijk niveau tekort schieten en bij beginners en halfgevoorderden op didactisch vlak. Dreyfus concludeert dat bestaande intelligente computersystemen, met name expertsystemen en intelligente onderwijssystemen, de gedachte ondersteunen dat de menselijke geest werkt zoals een computer. Zij bevorderen een opvatting van kennis als iets wat in expliciete regels en principes formuleerbaar moet zijn. Hierdoor raken de intuïtieve vaardigheden en expertise van mensen, die niet in formele regels zijn te vangen, gedevalueerd en worden leerlingen en studenten aangemoedigd om kennis en vaardigheden te verwerven volgens het rationalistische model. Uiteindelijk kan het zelfbeeld van mensen hier zo door veranderen dat zij zichzelf alleen nog beschouwen in rationalistische termen, als abstracte denkmachines. Het is deze tendens die Dreyfus wil keren.

Onderwijs op afstand

In *On the Internet* uit 2001 betoogt Dreyfus dat niet alleen ITS-en ongeschikt voor het bereiken van expertise zijn. Ook bij onderwijs op afstand – onderwijs waarbij de (menselijke) docent zich niet in dezelfde ruimte als de leerling bevindt, maar waar communicatie via het internet plaatsvindt – kan expertise niet bereikt worden. Maar blijft het steken op een niveau van competentie. Voor expertise is de fysieke aanwezigheid van een leraar vereist, want deze aanwezigheid maakt nauwkeurige en niet regelgebonden observatie van de leraar en betrokkenheid van de leerling mogelijk. Om een expert te worden is confrontatie met klasgenoten en een echte leraar noodzakelijk.

Ook zal, betoogt Dreyfus, onze betrokkenheid met de fysieke en sociale wereld afnemen als we het internet meer gebruiken. Hierdoor neemt onze realiteitszin en zinvolheid van ons leven af en trekken we ons steeds meer terug in een betekenisloze eenzame virtuele wereld. Dreyfus is niet onverdeeld negatief. Als we ons voor een doel inzetten, kan het internet onze krachten vergroten door relevante informatie te verstrekken en ons in contact met andere mensen met hetzelfde doel te brengen.

AI in 1965 voorspelde Dreyfus dat de symbolische AI grotendeels op een mislukking zou uitdraaien in haar streven naar een volledige imitatie van menselijke intelligentie. De voorspellingen en verwachtingen die aan nieuwe projecten en benaderingen binnen de symbolische AI zijn opgehangen heeft hij in de loop der jaren stelselmatig bekritiseerd. Het lijkt erop dat Dreyfus in veel opzichten gelijk heeft gekregen. Hoewel de symbolische AI zeker ook successen heeft geboekt, zijn de resultaten op veel terreinen teleurstellend. Zo zijn er nog geen computerprogramma's ontwikkeld die natuurlijke taal goed kunnen begrijpen, beelden foutloos kunnen interpreteren of creativiteit vereisende problemen kunnen oplossen. Niet alleen heeft Dreyfus gelijk gekregen in veel van zijn voorspellingen, AI-onderzoek is ook opgeschoven in de richting van Dreyfus' alternatieve theorie van intelligentie. Dit geldt voor de genoemde opkomst van de neurale netwerk-AI, waarvan de uitgangspunten, zoals Dreyfus zegt, goed verenigbaar zijn met zijn eigen ideeën over intelligentie.

De gesitueerdheid van intelligentie is een centraal uitgangspunt in het werk van de beroemde AI-onderzoeker Terry Winograd en zijn collega Fernando Flores. Zij willen niet alleen AI-onderzoek maar ook het ontwerpen van computersystemen stelen op Heideggeriaanse

uitgangspunten. Winograd en Flores stellen dat bij het ontwerpen van computersystemen in ogenschouw moet worden genomen dat deze systemen moeten functioneren in een menselijke wereld en moeten communiceren met menselijke gebruikers, en dat de interne logica van een computersysteem hierop afgestemd moet worden. Voorkomen moet worden dat computers hun eigen rationalistische logica opleggen aan de omgeving waarin ze functioneren.

Ook de idee dat intelligentie het hebben van een lichaam vooronderstelt heeft weerklank gevonden in AI-onderzoek. Dit wordt het best geïllustreerd door een recent project aan MIT, het internationaal veel aandacht trekkende Cog-project onder leiding van Rodney Brooks. Cog is een robot die door zijn sensomotorische interacties met de omgeving sensomotor-intelligentie moet verwerven, om op basis hiervan 'hogere' vormen van intelligentie te ontwikkelen.

Voor een niet onbelangrijk deel zijn deze ontwikkelingen terug te voeren op het werk van Dreyfus zelf. Dreyfus was degene die het ideeëngoed van denkers als Heidegger en Merleau-Ponty in de AI-wereld introduceerde. Het werk van AI-onderzoekers als Winograd en Flores en Agre en Chapman is heel expliciet op dit ideeëngoed geïnspireerd. Maar ook veel andere AI-onderzoekers, zelfs adepten van de symbolische AI zoals Marvin Minsky en John McCarthy geven toe dat de kritieken van Dreyfus invloed op hun onderzoek hebben gehad. Dreyfus heeft hiermee bewezen dat filosofen een belangrijke rol kunnen spelen als critici van, en commentatoren op zich ontwikkelende wetenschap en techniek.

Primaire Bibliografie

Dreyfus over Artificiële Intelligentie en computers

What computers can't do: a critique of artificial reason, (1972) Harper and Row, New York,

Mind over machine: the power of human intuition and expertise in the era of the computer,
(1986) Free Press, New York, met Stuart Dreyfus

What computers still can't do: a critique of artificial reason, (1992) MIT Press, Cambridge.

On the Internet, (2001) Routledge, London.

Overig werk

Being-in-the-world: a commentary on Heidegger's 'Being and Time', division I, (1991) MIT Press, Cambridge, Mass.

Michel Foucault: beyond structuralism and hermeneutics, (1982) Harvester, New York, met Paul Rabinow.