

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

Brey, P. (1997). 'Hubert Dreyfus - Mens versus Computer,' in *Van Stoommachine tot Cyborg*, Red. H. Achterhuis, Ambo, Baarn, 43-68.

Hubert Dreyfus: Mens versus Computer

1. Hubert Dreyfus, criticus van de artificiële intelligentie

In 1956, een kleine tien jaar na de uitvinding van de eerste programmeerbare digitale computers, werd op een congres in Dartmouth College in de Verenigde Staten de geboorte aangekondigd van een nieuw wetenschapsgebied, dat van de artificiële intelligentie. Artificiële intelligentie, of AI, zoals het vaak kort wordt genoemd, presenteerde zichzelf als een fundamentele wetenschap, die een systematische studie van het fenomeen 'intelligentie' nastreefde. Men wilde dit bereiken door intelligente processen te *simuleren* op computers. Uitgangspunt was dat het soort logische operaties die door computers worden uitgevoerd zo konden worden vormgegeven dat ze een gelijkenis vertoonden met menselijke denkprocessen. Men meende dat het in principe mogelijk was dat computers, indien juist geprogrammeerd, echte intelligentie bezaten, net zoals in kunstmatige omstandigheden vervaardigde diamanten desalniettemin echte diamanten zijn. Omdat de werkingen van een computer bekend zijn, in tegenstelling tot de werkingen van de menselijke geest, hoopte men zo tot wetenschappelijke verklaringen te komen van het verschijnsel intelligentie.

De AI was vanaf het begin een onderzoeksgebied met hoge doelen en grote beloften. Het hoogste doel was niets meer en niets minder dan de constructie van een computersysteem met de intelligentie en het redeneervermogen van een volwassen mens. Vele AI-onderzoekers meenden en voorspelden dat dit doel dankzij de uitvinding van de digitale computer en dankzij belangrijke ontwikkelingen op het gebied van de informatica en formele logica in slechts enkele decennia gerealiseerd kon worden. Zo voorspelde de beroemde AI-onderzoeker Herbert Simon in 1965 dat computers in 1985 in staat zouden zijn elke taak uit te voeren waartoe een mens ook in staat is. De evenzeer beroemde AI-onderzoeker Marvin Minsky voorspelde in 1967 dat alle belangrijke doelen van de AI binnen het tijdperk van een generatie zouden zijn gerealiseerd.

Dat dit soort voorspellingen indertijd serieus werd genomen valt te verklaren door de schijnbaar oneindige mogelijkheden die de computer leek te bieden. Hierbij speelden de vroege successen van de AI een belangrijke rol. Zo kon de AI al in het jaar van haar formele begin, 1956, het eerste succes aankondigen: een computerprogramma dat op beginnersniveau schaak kon spelen. Vrijwel ieder jaar daarna werden betere schaakprogramma's aangekondigd. Een programma uit 1964, STUDENT, was in staat korte stukjes tekst die algebra-problemen bevatten te interpreteren en deze problemen op te lossen. Een programma uit 1966, ELIZA, was in staat een bescheiden therapeutische dialoog aan te gaan met mensen over hun persoonlijke problemen. Financiers, zoals het Amerikaanse ministerie van defensie, waren overtuigd en jonge onderzoekers voelden zich aangetrokken tot deze nieuwe wetenschap. Dit was het begin van een grote groei van de AI, waarbij deze zich snel vestigde als een goed gefinancierd onderzoeksgebied, waarin wereldwijd honderden miljoenen guldens omgingen, en waarbinnen duizenden AI-onderzoekers werkzaam waren en zijn.

In het begin van de jaren zestig, toen de AI nog een jonge wetenschap was, kwam de ook nog jonge filosoof Hubert Dreyfus voor het eerst ermee in aanraking. Dreyfus was indertijd docent filosofie aan het wereldberoemde Massachusetts Institute of Technology (MIT). Zoals hij zelf verhaalt, gaf hij daar een keer een cursus over filosofische theorieën van waarnemen, kennen, en begrijpen. Zijn studenten vertelden

hem dat zulke filosofische theorieën inmiddels achterhaald waren door de opkomst van de computer. Aan MIT werd wetenschappelijk onderzoek gedaan onder leiding van Marvin Minsky, vertelden ze hem, dat als doel had een machine te construeren die al deze menselijke taken kon verrichten.

Geprikkeld door deze berichten zocht Hubert Dreyfus contact met zijn broer Stuart, op dat moment werkzaam als computerspecialist bij de RAND-organisatie, een breed georiënteerd onderzoeksinstituut. Via dit contact werd hij aangesteld als filosofisch consulent bij RAND, om hun nieuwe AI-onderzoek te evalueren. Dit werd geleid door twee AI-onderzoekers die later beroemd zouden worden met hun werk, Allen Newell en Herbert Simon. In de evaluatie van hun onderzoek kwam Dreyfus tot de conclusie dat hoewel het onderzoek aantoonde dat computers bepaalde problemen kunnen oplossen, het geen enkele substantiële bijdrage leverde aan een inzicht in het verschijnsel intelligentie, en volledig op de verkeerde weg was als poging om menselijke intelligentie te simuleren. Zijn vernietigende rapport *Alchemy and Artificial Intelligence*, geschreven in 1964, werd pas een jaar later door de RAND-organisatie vrijgegeven, nadat het aanvankelijk tegengehouden was door protesten van Newell en Simon. Het werd daarmee de eerste gedetailleerde gepubliceerde kritiek van de AI.

Het rapport van Dreyfus werd al snel door computerwetenschappers over de gehele wereld bediscussieerd. Het werd daarmee zijn eerste invloedrijke publikatie over dit onderwerp, het begin van een reeks boeken en artikelen met filosofische kritieken van de AI. Zijn belangrijkste publikatie op dit gebied, die hem internationale faam heeft bezorgd als criticus van de AI, is zijn boek *What Computers Can't Do* uit 1972. Nieuwe edities hiervan verschenen in 1979 en 1992. Ook van belang is een boek uit 1986, *Mind over Machine*, geschreven samen met zijn broer Stuart.

Wat de kritieken van Dreyfus zo bijzonder maakt is dat ze voor het grootste deel gemotiveerd zijn door een filosofische traditie die normaliter ver verwijderd is van de wereld van wetenschap en techniek. Dat is de traditie van de fenomenologie, zoals vertegenwoordigd door het werk van Martin Heidegger en Maurice Merleau-Ponty. De fenomenologie richt zich op het beschrijven van de manier waarop de mens zich verhoudt tot zijn wereld, en neemt hierbij de ervaring van deze mens zelf als uitgangspunt. Wanneer de fenomenologie uitspraken doet over wetenschap en techniek zijn deze meestal zeer algemeen en abstract. Fenomenologen zoals Heidegger en Merleau-Ponty doen echter wel meer concrete uitspraken over de aard van het menselijke waarnemen, denken en handelen. Hun ideeën hierover heeft Dreyfus op kundige wijze weten toe te passen in zijn kritiek van de AI.

Overigens is Dreyfus door de jaren heen naast zijn filosofische kritieken op de AI ook meer traditionele filosofische verhandelingen blijven afleveren, met name over het werk van fenomenologen als Heidegger, Merleau-Ponty, en Husserl, en andere filosofen als Foucault en Kierkegaard. Regelmatig terugkerend thema in deze publicaties is de wijze waarop de mens zijn wereld ervaart en manieren ontwikkelt om met deze wereld om te gaan. Zijn boek *Being-in-the-World*, uit 1991, wordt tegenwoordig gewaardeerd als een van de beste en helderste uiteenzettingen van het vroege werk van Heidegger.

Een regelmatig terugkerend thema in het werk van Dreyfus is zijn kritiek op het Cartesiaanse rationalisme. Uitgangspunten van het rationalisme, zoals geïnterpreteerd door Dreyfus, zijn dat de werkelijkheid een rationele structuur heeft (wetmatig is opgebouwd uit afzonderlijke elementen), dat het denken van de mens op dezelfde rationele manier werkt, en dat alles wat niet rationaliseerbaar is (niet in expliciete principes uit te drukken en te verdedigen is) weinig waarde heeft. Dreyfus is ervan overtuigd dat de Westerse cultuur nog steeds in sterke mate bepaald wordt door deze rationalistische opvattingen, maar meent, op basis van het werk van filosofen als Heidegger, Merleau-Ponty en Wittgenstein, dat deze fundamenteel onjuist zijn. Rationele, formele structuren zijn volgens hem menselijke constructies die achteraf op de werkelijkheid worden gedrukt. De kenbare werkelijkheid heeft in eerste instantie geen rationele structuur, maar een structuur die mede is bepaald door menselijke

behoefden en handelingen. De meest fundamentele manier van kennen is bovendien intuïtief, en niet rationeel. Het rationalisme, zoals het zich onder meer in de AI manifesteert, gaat aan deze oorspronkelijke structuur van de werkelijkheid voorbij, en doet onvoldoende recht aan intuïtieve kennis en vaardigheden, waardoor deze in de maatschappij gedevalueerd raken. Dreyfus is een constante pleitbezorger voor de intuïtieve kennis en vaardigheden van mensen, en een criticus van het rationalisme in zijn huidige manifestaties.

De invloed van Dreyfus' kritieken van de AI is in veel opzichten aanzienlijk geweest. Dreyfus heeft naam gemaakt onder AI-onderzoekers en in AI geïnteresseerde filosofen, hoewel aanvankelijk vooral in negatieve zin, omdat men zijn kritieken onsympathiek en weinig overtuigend vond. Het komt op zich al weinig voor dat het werk van een filosoof op grote schaal wordt bestudeerd door vakwetenschappers buiten de filosofie. Nog bijzonderder is echter dat verschillende invloedrijke AI-onderzoekers Dreyfus' kritieken serieus ter harte hebben genomen, en zijn alternatieve, op de fenomenologie gestoelde, ideeën over intelligentie verder zijn gaan uitwerken en toepassen in hun eigen onderzoek. Op deze manier heeft het vaak abstracte filosofische ideeëngoed waarop Dreyfus zich beroept een directe invloed gehad op de ontwikkeling van de AI.

Allereerst zal nu een schets gegeven worden van het klassieke AI-onderzoek. Vervolgens zullen de kritiek van Dreyfus op de klassieke AI en zijn alternatieve theorie van intelligentie worden uiteengezet. Daarna wordt Dreyfus' kritiek op een belangrijke recente benadering in de AI besproken, die van neurale netwerken. In de daarop volgende sectie wordt Dreyfus' kritiek op maatschappelijke toepassingen van intelligente computersystemen behandeld. In een concluderende sectie zal zowel de invloed als het gelijk van Dreyfus worden geëvalueerd.

2. Het klassieke paradigma van de artificiële intelligentie als nieuwste vertegenwoordiger van het rationalisme

Vanaf het ontstaan van de AI zijn er veel verschillende soorten van AI-onderzoek geweest, die gebruik maken van verschillende doelstellingen, methoden, en formalismen. Toch zijn vanaf het ontstaan van AI-onderzoek in de jaren vijftig tot en met het begin van de jaren tachtig de gedeelde eigenschappen van de verschillende soorten van AI-onderzoek zo overheersend dat van een paradigma kan worden gesproken, in de zin van de wetenschapsfilosoof Thomas Kuhn: een door onderzoekers gedeelde verzameling van methoden, doelen, aannamen, en exemplarische voorbeelden van succesvol onderzoek, die gezamenlijk een onderzoeksprogramma definiëren. Dit paradigma, dat ook nu nog een deel van het AI-onderzoek karakteriseert, is bekend onder verschillende namen. Het zal hier worden omschreven, om redenen die genoemd zullen worden, als 'symbolische AI,' of soms ook 'klassieke AI'.¹

De symbolische AI had in de eerste decennia van haar bestaan als doel om intelligente computersystemen te maken, met als hoogste inzet een systeem dat universele intelligentie bezat, dat wil zeggen: universele vermogens om te redeneren, problemen op te lossen, taal te begrijpen, en andere intelligente taken te verrichten, zoals een intelligent volwassen mens dit ook kan. Dit onderzoek was in eerste instantie nauwelijks gericht op technische toepassingen, en profileerde zich vooral als

¹ De veronderstellingen van de klassieke AI over de aard van intelligentie zijn niet alleen binnen dit vakgebied terug te vinden, maar hebben sinds de zeventiger jaren geleid tot een nieuw, interdisciplinair vakgebied, dat van de cognitiewetenschap ('cognitive science'). Cognitiewetenschap is een interdisciplinaire wetenschap van (biologische en kunstmatige) intelligente processen, die is ontstaan door samenwerkingen tussen AI-onderzoekers, psychologen, filosofen, en taalwetenschappers. Tegenwoordig beroept nog maar een deel van het onderzoek in de cognitiewetenschap zich op de uitgangspunten van de klassieke AI, en hebben zich ook andere onderzoekstradities gevormd, zoals het later te bespreken paradigma van neurale netwerken.

wetenschap: de nieuwe wetenschap van intelligentie. Sommige klassieke AI-onderzoekers, zoals Newell en Simon, stelden zich hierbij als expliciet doel van hun onderzoek het modelleren van cognitieve processen (denkprocessen) van mensen. Werk met dit expliciete doel wordt soms Cognitieve Simulatie genoemd. AI-programma's worden hierbij gezien als simulaties *en* verklaringen van menselijk intelligent gedrag.

Andere onderzoekers in de symbolische AI, zoals Marvin Minsky, pretendeerden niet dat hun computerprogramma's menselijke denkprocessen simuleerden, maar wel dat hun werk een theoretische bijdrage leverde aan het begrijpen van het verschijnsel intelligentie, door algemene eigenschappen van intelligente processen bloot te leggen. Zij stelden dat hun onderzoek misschien niet direct inzicht verleende in de psychologische uitvoering van intelligente taken, maar wel in de menselijke competentie in het vertonen van intelligent gedrag: het geeft een algemeen inzicht in het soort cognitieve vermogens die mensen moeten bezitten om intelligent gedrag te kunnen vertonen. De verschillen tussen de Cognitieve Simulatie-benadering en deze meer gebruikelijke benadering zijn uiteindelijk echter minder van belang dan hun overeenkomsten: beide benaderingen stellen zich tot doel het verschijnsel intelligentie te begrijpen, en beide benaderingen delen dezelfde theoretische uitgangspunten, methoden, en formalismen, namelijk die van de symbolische AI.

De naam 'symbolische AI' is ontleend aan het centrale uitgangspunt ervan, namelijk dat intelligentie een kwestie is van het verwerken en manipuleren van symbolen volgens vaststaande, formele regels. In wat volgt zal geprobeerd worden dit uitgangspunt plausibel te maken. Een eerste stap op weg ernaartoe is de aanname dat intelligente processen, zoals waarnemen, redeneren, rekenen, en taal gebruiken, met elkaar gemeen hebben dat ze vormen van *informatieverwerking* zijn. Informatieverwerken bestaat uit het opnemen van informatie uit de omgeving, het verwerken of manipuleren van deze informatie, en het geven van een respons. Zo houdt het maken van sommen in dat men zich eerst informeert wat de som is, vervolgens op grond van deze informatie denkoperaties uitvoert, en vervolgens tot een uitkomst komt. Het spelen van schaak vereist dat men de bordsituatie in zich opneemt, nadenkt, en tenslotte tot een zet komt. Het lijkt kortom een veronderstelling dat wat intelligente organismen en systemen gemeen hebben is dat ze *informatieverwerkende systemen* zijn.

Als men deze aanname heeft gemaakt, kan men zich vervolgens afvragen: wat is bij intelligente systemen het karakter van deze informatie, en op welke manier wordt deze informatie 'verwerkt'? De symbolische AI doet hierover twee aannamen. Ten eerste wordt aangenomen dat informatie om door een informatieverwerkend systeem verwerkt te kunnen worden *gerepresenteerd* moet worden. Om informatie in zich op te nemen moet het systeem beschikken over een medium waarop deze informatie kan worden ingeschreven. Een dergelijk medium, dat informatie verschaft over een werkelijkheid buiten zichzelf, wordt een *representatie* genoemd. Alledaagse voorbeelden van representaties zijn foto's, schilderijen, grafieken, en geschreven en gesproken zinnen. Dit is echter niet het soort representaties dat door informatieverwerkende systemen kan worden gebruikt. Zulke 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 innerlijke, mentale representaties, waarop gedachten, herinneringen en waarnemingen zijn ingeschreven.

De volgende aanname van de klassieke AI is de meest kenmerkende: Zij stelt dat de innerlijke representaties van intelligente systemen *symbolisch* van aard zijn. Men had bijvoorbeeld ook, met verwijzing naar representaties als foto's en schilderijen, kunnen aannemen dat innerlijke representaties *iconisch* zijn: dat ze informatie dragen doordat ze een uiterlijke gelijkenis bezitten met datgene waarnaar ze verwijzen, zoals een schilderij informatie verschaft over het geportretteerde door erop te lijken. Maar de klassieke AI

gaat ervan uit dat innerlijke representaties meer overeenkomen met natuurlijke taal. Taal is symbolisch, dat wil zeggen, taaltokens hebben geen enkele gelijkenis 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.

De aanname dat interne representaties, als de dragers van informatie in intelligente systemen, symbolisch zijn, lijkt beter dan dat ze bijvoorbeeld iconisch zijn. Zo zijn iconische representaties niet in staat om abstracte zaken 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 oneindig aantal verschillende informatie-inhoudende representaties representeren. Dit kan men zien in de natuurlijke taal, waar een eindig aantal woorden kan worden gecombineerd tot een in principe oneindig aantal verschillende zinnen. Bovendien lijken symbolen in het soort cognitieve taken die soms als de meest intelligente worden gezien, zoals het maken van wiskundige berekeningen en logisch redeneren, een prominente rol te spelen.

Uit de eerdere aanname dat intelligentie berust op de capaciteit om informatie te verwerken, en de vervolgens beredeneerde aanname dat informatieverwerken bestaat uit het manipuleren van symbolen, volgt nu dat intelligente systemen *symboolverwerkende* systemen zijn. Er is tot nu toe echter nog geen aanname gemaakt over de manier waarop deze symbolen worden verwerkt. De aanname van de symbolische AI is hier dat symbolen uitsluitend worden 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.

Maar hoe wordt bepaald welke verwerkingen er door het systeem worden uitgevoerd op basis van deze formele eigenschappen? Hier volgt een andere fundamentele aanname: een informatieverwerkend systeem beschikt over regels volgens welke deze symbolen geïnterpreteerd en verwerkt worden. Dit zijn *formele* regels, omdat ze slechts aangrijpen op de formele eigenschappen van deze symbolen. Ze werken als het ware automatisch: in respons op een aangeboden symbool of reeks symbolen voeren ze een verwerking uit, waaruit weer nieuwe symbolen(reeksen) resulteren, die vervolgens weer automatisch worden aangegrepen door een andere regel, etc. Als zulke regels niet zouden bestaan, dan lijkt het dat intelligentie een mysterie moet blijven, omdat dan niet makkelijk verklaard kan worden hoe symbolen in een intelligent systeem verwerkt worden.

Ook lijkt de aanname dat intelligente informatieverwerking het toepassen van regels inhoudt plausibel, omdat 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, zo lijkt logisch redeneren het toepassen van regels van de logica in te houden, en lijkt het oplossen van wiskundige of natuurkundige problemen het toepassen van wiskundige stellingen of natuurkundige wetten 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 hierboven 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 ook nog 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.

De opvattingen over intelligentie van de symbolische AI worden soms als vernieuwend beschouwd, maar Dreyfus benadrukt dat het hier slechts gaat om de nieuwste reïncarnatie van een oude, in de geschiedenis van het denken steeds weer terugkerende theorie van menselijke intelligentie, die met name gevonden kan worden in wat soms de *rationalistische* traditie in de filosofie wordt genoemd. Plato heeft de basis gelegd voor deze theorie. Hij stelde dat wijsheid bestond uit het kunnen formuleren van kennis in expliciete definities, en had een afkeer van mensen die handelden op basis van aangeleerde vaardigheden of intuïtie. 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.

De belangrijkste representant van dergelijke ideeën uit het moderne tijdperk is René Descartes, die in de zeventiende eeuw beweerde dat elk probleem in afzonderlijke basiselementen is op te splitsen, en dat elke complexe werkelijkheid of gedachte te verklaren is door een systeem van regels waarmee het uit dergelijke basiselementen gevormd kan worden. Hij meende dat ook de menselijke geest volgens dergelijke regels en basiselementen werkzaam is. Na Descartes vindt men dergelijke opvattingen over intelligentie terug bij veel andere rationalistisch ingestelde denkers, zoals 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.

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 veronderstelling*, 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 deze 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 eerder genoemde Cognitieve Simulatie-benadering. In alle variëteiten van symbolische AI wordt echter wel een tweede aanname gedaan, de *epistemologische veronderstelling* 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.

Zowel de epistemologische als de psychologische veronderstelling zijn tenslotte vaak mede gebaseerd op de *ontologische veronderstelling*, dat de door de mens kenbare werkelijkheid een formaliseerbare structuur heeft, 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 deze zowel in de epistemologische als psychologische veronderstelling worden verondersteld.

Dat de symbolische AI een rationalistische opvatting van intelligentie incorporeert die al eerder in de geschiedenis van het denken is te signaleren, is op zich

misschien niet zo interessant, ware het niet dat er in de loop van de geschiedenis belangrijke filosofische kritieken op een dergelijke visie zijn verschenen. 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.

3. Dreyfus' kritiek op de symbolische AI

De belangrijkste twee wetenschapsgebieden die bewijs zouden kunnen aanleveren voor de psychologische veronderstelling van de symbolische AI zijn de psychologie en de neurofysiologie. De neurofysiologie wordt van belang gevonden omdat denken met behulp van symbolen en regels alleen mogelijk wordt geacht als deze symbolen en regels geïmplementeerd zijn in de menselijke hersenen, net zoals een computerprogramma geïmplementeerd is in de hardware van een computer. De veronderstelling dat in de hersenen een symboolverwerkend systeem is geïmplementeerd noemt Dreyfus ook wel de *biologische veronderstelling*, een vierde veronderstelling die in de symbolische AI vaak voorkomt. Dreyfus betoogt echter dat onderzoek in zowel de psychologie als de hersenwetenschappen geen goed empirisch bewijs heeft kunnen verschaffen voor de psychologische en biologische veronderstellingen. Tegelijkertijd bestaat er ook geen afdoende bewijs voor de onjuistheid van deze veronderstellingen.

Dreyfus' belangrijkste kritiek is echter gericht op de epistemologische veronderstelling, terug te vinden in alle vormen van symbolische AI, dat intelligent gedrag reproduceerbaar is door menselijke kennis te formaliseren (in regels uit te drukken) en in een computer op te slaan. 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 *reproducen*. Het hebben van kennis vooronderstelt volgens Dreyfus de vaardigheid om deze kennis in relevante situaties te kunnen reproducen 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. Gebeurt dit niet, dan is deze kennis niet werkelijk aanwezig.

Het toepassen van geformaliseerde, in regels weergegeven kennis lijkt hier echter 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. Dit lijkt slechts een kwestie te zijn van het vergelijken van de vorm van het nieuw gegeven symbool met de vorm van in het programma aanwezige symbolen. Vaak ligt de zaak echter niet zo eenvoudig. Stel bijvoorbeeld dat men een computer probeert te programmeren om taal te begrijpen. De meest eenvoudige methode hiervoor zou zijn om de computer uit te rusten met een aantal interpretatieregels die voor afzonderlijke woorden een definitie van hun betekenis geven en een aantal grammaticaregels om zinnen te kunnen ontleden. 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.

Eén van de vele problemen die hierbij echter optreden, is het feit dat sommige woorden dubbelzinnig zijn. Als de computer het woord 'heet' krijgt aangeboden, zal het meestal de interpretatieregels 'Als iets heet is, heeft het een hoge temperatuur' kunnen toepassen, maar als het gaat om hete gerechten kan de interpretatieregels '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, bijvoorbeeld 'Als in voorafgaande tekst wordt verwezen naar gepeperde gerechten, pas dan de tweede interpretatieregels toe.' Ook op deze toepassingsregels zullen echter weer uitzonderingen bestaan. Zo kan een tekst gaan over hete Mexicaanse gerechten, waarbij echter duidelijk uit de context blijkt dat op de hoge temperatuur ervan wordt gedoeld. Er zijn dus ook uitzonderingen op de toepassingsregels, en daarom zijn er ook weer toepassingsregels nodig voor het goed toepassen van de toepassingsregels. Zo dreigt er een oneindige regressie van regels te ontstaan, waardoor interpretatie onmogelijk wordt gemaakt.

Samengevat lijkt het probleem te zijn dat de juiste interpretatie van veel gegevens sterk afhankelijk is van allerlei omliggende gegevens. Formele regels zijn er echter op gericht zo min mogelijk acht te slaan op context en slechts oog te hebben voor het gegeven of de paar 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.

Dreyfus observeert dat mensen wél moeiteloos gegevens interpreteren vanuit hun context. Zo merken mensen het vaak niet eens als in een tekst een woord verkeerd is gespeld en vullen zij vanuit de context automatisch de goede betekenis in, terwijl computers dan meestal hopeloos vastlopen. 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 verstand-probleem* van de symbolische AI. Hij meent echter dat, om bovengenoemde redenen, dit probleem onoplosbaar is.

Computers functioneren het beste wanneer de 'wereld' die zij moeten tegemoet treden en interpreteren een kunstmatige, formele wereld is. Een formele wereld bestaat uit elementen waarvan de identiteit direct afgelezen kan worden uit de vorm ervan, onafhankelijk van andere elementen in die wereld en waarin elementen volgens duidelijke wetmatigheden met elkaar in verband staan. In een dergelijke geformaliseerde werkelijkheid doet het gezond verstand-probleem zich nauwelijks voor. Dit geldt bijvoorbeeld voor spelen als schaak en boter-kaas-en-eieren, voor wiskunde, en voor formele logica. In zulke domeinen gaan geformuleerde problemen bovendien meestal vergezeld van een duidelijk doel of 'eindtoestand' die bereikt moet worden, bijvoorbeeld het hebben van drie kruisjes op een rij, het mat zetten van de koning, of een getalswaarde vinden bij een wiskundige vergelijking. En er zijn duidelijke regels volgens welke stappen genomen kunnen worden om het doel te bereiken. Bij het oplossen van dit soort problemen heeft de symbolische AI haar grootste successen geboekt.

Meer alledaagse probleemsituaties waarin intelligent gedrag vereist is, lijken echter een heel andere structuur te hebben dan deze kunstmatige 'werelden'. Neem bijvoorbeeld een alledaags probleem zoals het per ongeluk in de auto sluiten van de autosleutels bij het bezoeken van de supermarkt. Hoewel duidelijk is dat deze situatie een probleem oplevert die een oplossing behoeft, hoeft ten eerste het doel van deze oplossing niet bij voorbaat duidelijk te zijn. Is het doel om weer te beschikken over de autosleutels in de auto? Niet als men gemakkelijker kan beschikken over een stel reservesleutels die bijvoorbeeld een partner even kan komen brengen. Is het doel om weer te kunnen rijden in de auto? Niet als het een belangrijkere prioriteit is om vroeg thuis te zijn. Het zoeken naar een oplossing is niet gericht op één uniek doel, maar is een continue afweging van verschillende behoeften, zoals het beperken van schade aan de auto, van kosten of tijdverlies, het bijtijds thuiskomen, etc.

Ten tweede 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. Er lijkt niet een vast aantal voorwerpen met objectieve eigenschappen en relaties aanwezig te zijn, waar vervolgens regels op toegepast kunnen worden. Pas tijdens het oplossen van het probleem komen potentieel relevante feiten in beeld, zoals een autoraampje dat op een kier staat of een op de grond slingerende metalen strip waarmee men kan inbreken of een vorige eigenaar die misschien nog een stel reservesleutels heeft. Het oplossen van het probleem zal typisch een aantal stadia doorlopen, waarin men het op verschillende manieren probeert te conceptualiseren, om zo een voorstelling van het probleem te vinden die het meeste gevoel geeft dat men greep op de situatie kan krijgen. Dit creatief kunnen herformuleren van het probleem lijkt een meer essentiële vaardigheid te zijn dan het vinden van een oplossing wanneer een goede probleemdefinitie reeds gevonden is. Formele regels lijken in het zoeken naar een goede probleemdefinitie geen rol te spelen.

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.

4. Intelligentie is belichaamd en gesitueerd

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. 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, die vooral stoelt op de filosofie van Heidegger (en in mindere mate op die van Wittgenstein en Merleau-Ponty) is waarschijnlijk het moeilijkst te begrijpen onderdeel uit het werk van Dreyfus.

Dat de psychologische veronderstelling, dat mensen representaties en regels nodig hebben om de wereld te interpreteren, zo redelijk lijkt, komt volgens Dreyfus omdat er vaak wordt uitgegaan van een bepaalde opvatting van wat de wereld is, en hoe deze door mensen wordt gekend. De wereld wordt dan geïnterpreteerd als een materiële, van de mens onafhankelijke structuur zoals die ons door de natuurwetenschappen wordt voorgespiegeld. Deze wereld is inherent betekenisloos, en is ruimtelijk gescheiden van de mens, zodat geen directe, ongemedieerde ervaring ervan mogelijk is. Dreyfus ontkent niet de waarde van het door de natuurwetenschappen aangeleverde perspectief over wat verstaan moet worden onder de 'wereld', 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. Onder 'wereld' wordt dan verstaan de wereld zoals die zich manifesteert in de menselijke ervaring.

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, dat de mens actief is en leert waarnemen, handelen en denken. Deze wereld evolueert echter mee met deze activiteiten, omdat mede in en door deze activiteiten zich steeds weer nieuwe structuren manifesteren. 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.

Dreyfus betoogt dat bij mensen de ervaring van de wereld in zijn geheel altijd voorafgaat aan een ervaring van afzonderlijk te onderscheiden elementen. Zo ervaart een depressief persoon de wereld als 'grauw' en 'zinloos' nog voordat specifieke elementen in de wereld onderscheiden worden en wordt een nieuwe omgeving als 'veilig' of 'bedreigend' ervaren nog voordat discrete objecten in deze omgeving worden onderscheiden; het is de situatie in zijn geheel die deze ervaring oproept.

Specifieke elementen in een wereld of situatie worden onderscheiden en ervaren vanuit deze meer algemene ervaring van betekenis en zin. Hierdoor staan deze elementen altijd in een zinvol betekenisverband met hun context. Zo ervaart een timmerman tijdens zijn werk een rondslingerende hamer als een 'ding-om-mee-te-hameren' en als 'ding-dat-gebruikt-wordt-met-spijkers', maar zal hij de hamer in een meer bedreigende betekeniscontext misschien eerder ervaren als een 'wapen-tegen-een-indringer'. In geen van beide gevallen zal hij de hamer waarnemen als een van de omgeving onafhankelijk voorwerp waarvan de betekenis en zin nog vastgesteld moeten worden. Op dezelfde manier 'ziet' een schaakgrootmeester een betekenisvol bordpatroon en de bijbehorende goede zet die door dit patroon wordt gesuggereerd, zonder dat hij in dit patroon afzonderlijke schaakstukken hoeft te ontwaren en de voor deze stukken geldende spelregels hoeft toe te passen.

Intelligent gedrag vereist dat een mens in de situatie waarin hij of zij zich bevindt een betekenisvolle structuur ontwaart die handelingen voorschrijft welke gegeven de situatie zinvol zijn. De betekenisvolle structuur die uiteindelijk waargenomen wordt, is het lokale produkt 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 volgen handelingen als automatisch; zoals het oog automatisch de invallende hoeveelheid licht 'begrijpt' en zich aanpast door de pupil te vergroten of te verkleinen, 'begrijpt' de mens de situatie waarin hij zich bevindt en reageert hij vanuit dit begrip met een bijbehorend, ter plekke opgebouwd handelingspatroon. 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. Dat zou impliceren dat de veronderstellingen van de symbolische AI juist zijn. 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. Voor intelligent gedrag is het echter meestal niet nodig om op deze manier te abstraheren, behalve voor probleemsituaties die zelf al abstract zijn gedefinieerd.

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 opvatting, die niet geheel losstaat van de eerste, is vooral geïnspireerd op de filosofie van Merleau-Ponty. Dreyfus' uiteenzetting van deze opvatting is echter vaak, net zoals 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 ook kan bestaan zonder een lichaam, maar alleen ontwikkeld kan worden met behulp van een lichaam. Beide opvattingen zullen hier op hun plausibiliteit beoordeeld worden.

(1) *Vereist intelligentie het hebben van een lichaam?* Zowel door wetenschappers als

leken wordt vaak aangenomen dat intelligentie gelokaliseerd is in de hersenen. Voor tenminste één belangrijk type intelligentie, *sensomotorische intelligentie*, is deze aanname echter omstrepen. Sensomotorische intelligentie is de vaardigheid die mensen hebben in waarnemen, herkennen, bewegen en manipuleren en het coördineren en integreren van waarneming en beweging. Het ontwikkelen van sensomotorische intelligentie vereist duidelijk een lichaam, maar dit bewijst op zich nog niet dat sensomotorische intelligentie mede in het lichaam gelokaliseerd raakt. Het is in principe mogelijk dat sensomotorische intelligentie uitsluitend een vaardigheid van de hersenen is in het interpreteren van door de zintuigen doorgegeven prikkels en het aansturen van het spierstelsel.

Een alternatieve, even goed verdedigbare hypothese is echter dat sensomotorische intelligentie gelokaliseerd is in een complex feedback-systeem dat zowel de hersenen, de rest van het zenuwstelsel, de zintuigen en de spieren (en hormonale klieren) omhelst. Zintuigen en spieren kunnen dan even goed beschouwd worden als informatieverwerkende systemen, of onderdelen daarvan, als de hersenen. Sensomotorische intelligentie is dan een eigenschap van een zich ontwikkelend lichaam, waarin niet alleen de hersenen, maar ook andere organen een trainingsproces doormaken waardoor zich een totaalsysteem ontwikkelt dat in staat is tot intelligent, gecoördineerd waarnemen en bewegen.

Ook als deze hypothese correct is, is het echter onwaarschijnlijk dat alle intelligentie van de mens in het lichaam gedistribueerd is. Met name abstracte, 'hogere' vormen van intelligentie, zoals abstract redeneren en rekenen, lijken niet afhankelijk te zijn van het lichaam. Mensen kunnen ledematen en organen kwijtraken en verlamd raken en toch hun abstract denkvermogen behouden, en het lijkt tenminste een theoretische mogelijkheid, zoals dat wel eens gebeurt in science fiction verhalen, dat iemands hersenen worden verwijderd uit zijn lichaam en onder kunstmatige omstandigheden voortleven, met behoud van het abstracte denkvermogen van deze persoon. Niet alle typen intelligentie lijken dus het hebben van een lichaam te vereisen.

(2) *Kan intelligentie alleen ontwikkeld worden met behulp van een lichaam?* Zelfs als het hebben van een lichaam niet vereist is voor het *bezitten* van intelligentie, kan het nog wel vereist zijn voor het *ontwikkelen* van intelligentie. Dat een lichaam vereist is voor het ontwikkelen van sensomotorische intelligentie spreekt vanzelf. Maar voor meer abstracte vormen van intelligentie is dit uitgangspunt minder plausibel. Een alternatieve hypothese, verenigbaar met de psychologische veronderstelling van de symbolische AI, is dat abstracte intelligentie berust op een aangeboren symboolsysteem in de hersenen, dat zich in principe onafhankelijk van het lichaam kan ontwikkelen, net zoals een computer geen lichaam nodig heeft om zijn kennisbestand uit te breiden.

De eerste abstracte denkvermogens die kinderen ontwikkelen lijken echter nog sterk geïntegreerd te zijn met hun sensomotorische intelligentie. Zo is hun eerste gebruik van taal sterk gebonden aan de wereld waarin ze waarnemen en handelen, en zijn hun eerste rekenoefeningen gerelateerd aan concrete voorwerpen. Ook hun voorstellingsvermogen betreft in eerste instantie alleen deze sensomotorische wereld. Een alternatieve hypothese voor de ontwikkeling van abstracte intelligentie is nu dat abstracte intelligentie niet berust op fundamenteel nieuwe vaardigheden, maar direct voortbouwt op de vaardigheden die reeds zijn ontwikkeld in de ontwikkeling van sensomotorische intelligentie.

Sensomotorische intelligentie omvat vaardigheden als patroonherkennen, het visueel scannen van beelden, het mentaal groeperen en manueel manipuleren van voorwerpen, het inschatten van effecten van krachten, het visueel opdelen en transformeren van ruimtelijke structuren en het mentaal anticiperen van de effecten van handelingen. Een zich ontwikkelende abstracte intelligentie zou direct kunnen voortbouwen op deze vaardigheden door ze toe te passen in meer abstracte domeinen. Zo zou zelfs het manipuleren van abstracte symbolen, zoals in wiskunde en formele logica, uiteindelijk terug te voeren kunnen zijn op ons vermogen om materiële

voorwerpen te manipuleren. Dreyfus lijkt deze opvatting tegenwoordig ook te huldigen, en verwijst in zijn recente werk naar studies van Mark Johnson (1987), 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 boven 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.

5. Het nieuwe paradigma van neurale netwerken

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. Neurale netwerken nemen afstand van de idee dat intelligent gedrag voortkomt uit het manipuleren van symbolen volgens formele regels. De inspiratiebron voor het modelleren van intelligente processen wordt niet meer gevonden in de digitale computer, maar in de structuur en werking van de menselijke hersenen. Wat neurale netwerken-AI nog wel gemeen heeft met symbolische AI is dat intelligentie wordt gezien iets dat berust op het verwerken van informatie.

Neurale netwerken zijn geïnspireerd op de structuur en werking van het menselijke zenuwstelsel, met name de hersenen. Het zenuwstelsel is opgebouwd uit zenuwcellen (neuronen). Zenuwcellen zijn op te vatten als informatieverwerkende systeempjes: ze ontvangen prikkels, van andere zenuwcellen of soms direct van zintuigcellen waarmee ze in contact staan en reageren hierop door zelf elektrochemisch opgewekte prikkels af te geven aan andere zenuwcellen, of soms ook spieren en klieren, waar ze op aangrijpen via zenuwuitlopers. Of een zenuwcel zelf prikkels afgeeft en hoe sterk deze zijn, wordt bepaald door een fysiologisch bepaald '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. Zenuwcellen zijn dus te begrijpen als verwerkingseenheden (processors) met tamelijk eenvoudige input/output functies.

Men meent nu dat het verschil tussen menselijke zenuwstelsels met lage en met hoge intelligentie voornamelijk wordt bepaald door de manier waarop de zenuwcellen hierin met elkaar en met de rest van het lichaam verbonden zijn. Intelligentie is dus voornamelijk een produkt van de connecties die zenuwcellen aangaan (vandaar de naam *connectionisme*). Zenuwcellen 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 zenuwcellen aangaan nog tamelijk willekeurig, maar naarmate meer geïnterageerd wordt met zijn of haar omgeving wijzigen de connecties van zijn of haar 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.

De neurale netwerken-AI probeert artificiële intelligentie te creëren door op basis van de werking van het zenuwstelsel modellen te bouwen die bestaan uit netwerken van eenvoudige verwerkingseenheden met input/output functies die lijken op die van zenuwcellen. Neurale netwerken bestaan uit een aantal processors dat meestal varieert van enkele tientallen tot enkele duizenden, die verbindingen met elkaar onderhouden, en de sterkte van hun verbindingen wijzigen op basis van door henzelf ontvangen

prikkels. Ze bestaan typisch uit een input-laag van neurale eenheden waar informatie wordt ingevoerd, een of meer tussenlagen en een output-laag waar informatie wordt uitgevoerd. In de praktijk worden neurale netwerken meestal niet echt fysisch nagebouwd maar gesimuleerd op gewone, digitale computers. De laatste jaren is er echter een opkomst van parallelle computers die soms tienduizenden parallel werkende processors bezitten en werkt men ook aan 'analoge' parallelle computers gebaseerd op glasvezeltechnologie.

Bestaande 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 en categorisatie. 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.²

Hubert Dreyfus stelt dat de uitgangspunten van neurale netwerken-AI goed verenigbaar zijn met zijn eigen visie op intelligentie (Dreyfus & Dreyfus, 1988; Dreyfus 1992). Neurale netwerken stappen af van de 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 (uiteindelijk) 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 dit hoge streven waar te maken. Dit ligt niet aan de uitgangspunten van de neurale netwerken-AI maar aan de grote complexiteit van menselijke intelligentie. Het fundamentele probleem met neurale netwerken-AI is dat ook hier weer het eerder genoemde gezond verstand-probleem opdoemt, zij het in een iets andere vorm. De intelligentie van neurale netwerken berust grotendeels op ervaring: de in het verleden gekweekte verbindingen zijn tot stand gekomen opdat het netwerk goed zou kunnen omgaan met situaties die zich in het verleden voordeden. Hun vermogen om met nieuwe situaties intelligent te kunnen omgaan berust op hun vaardigheid in het intelligent *generaliseren* van oude naar nieuwe situaties. Het lijkt er echter op dat intelligent generaliseren goede criteria vereist voor hoe gegeneraliseerd moet worden: welke oude kennis is relevant in de nieuwe situatie en welke aanpassingen zijn nodig in deze kennis wanneer hij wordt toegepast op de nieuwe situatie? Als je in het verleden bijvoorbeeld alleen binnenshuis appels hebt gegeten en je wilt nu een appel buitenshuis eten, mag je dan vanuit je ervaring in het verleden wel concluderen dat appels ook buitenshuis eetbaar zijn?

Allerlei kennis waarover een mens beschikt kan relevant zijn bij het generaliseren naar nieuwe situaties. Men kan niet van tevoren weten welke gegevens wel en welke niet relevant zijn. In potentie moet een mens dus de beschikking hebben over alle kennis die in zijn hersenen opgeslagen is om succesvol te kunnen generaliseren. Deze eis geldt dan echter ook voor het succesvol generaliseren van neurale netwerken. Dit

² Zie Meijering (1993) voor een korte, en Anderson (1995) voor een uitgebreidere inleiding tot neurale netwerken.

zou betekenen dat een netwerk dat succesvol kan generaliseren net zoals de hersenen zou moeten bestaan uit miljarden processors, 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 science fiction.

6. Van medisch specialist tot leraar: intelligente computersystemen in de maatschappij

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. Het uiteindelijke doel van het meeste recente AI-onderzoek is om te komen tot interessante technische toepassingen. AI-onderzoek heeft meestal niet meer de pretentie, zoals in de Cognitieve Simulatie-benadering, om een fundamentele bijdrage leveren aan het wetenschappelijk verklaren van intelligentie. Veel AI-onderzoekers omschrijven zichzelf dan ook niet meer als wetenschappers, maar bijvoorbeeld als 'taaltechnologen' of 'kennisingenieurs'.

AI als technologie is 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 besturingssystemen 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. Intelligente computersystemen maken *keuzen* en nemen *beslissingen* volgens criteria waarop de gebruikers van deze systemen weinig of geen greep hebben. Zij nemen dus keuze- en beslissingsverantwoordelijkheden van mensen over. Het overlaten aan computersystemen van sommige keuzen, zoals de belichtingskeuzen die worden gemaakt door een 'intelligente' videocamera of de tegenzetten die worden gedaan door een schaakcomputer, lijkt vanuit ethisch oogpunt gezien weinig problematisch. Maar bij andere beslissingen, zoals de beslissing of een asielzoeker wel of niet wordt toegelaten, of welke strafmaat een wetsovertreder krijgt, is het overlaten ervan aan een computer vanuit ethisch oogpunt zeer omstreden.

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 diagnostiseren 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.

In zijn vroege werk was 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 en doctoren in de natuurkunde, lijkt vaak terug te voeren op aangeleerde formele regels, zoals de regels van het schaakspel, natuurwetten en wiskundige principes, 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 Dreyfus 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 halfgevorderde in een bepaald kennisdomein, een beginstructuur die een versimpelde kijk op de werkelijkheid geeft, maar een handvat biedt vanwaaruit de eigenschappen en vereisten van specifieke gevallen aangeleerd kunnen worden. Omdat (*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 het niveau van echte expertise nooit halen.³

³ Een uitzondering geldt hier voor kennisdomeinen die wel geheel formaliseerbaar zijn, zoals sommige domeinen van de wiskunde en formele logica. Schaak is een grensgeval; de spelregels zijn formaliseerbaar, en daarom zijn schaakcomputers op expert-niveau in principe mogelijk, zoals de schaakcomputer Deep Blue heeft bewezen door in mei 1997 een schaakmatch van wereldkampioen Kasparov te winnen. Deep Blue heeft echter slechts kunnen winnen doordat zijn eigen strategie uitvoerig is afgestemd op de in het verleden vertoonde speelstijl van Kasparov. Er zijn geen 'beste zetten' bij schaak; of een zet de winstkansen optimaliseert is

Hiermee komt een aantal beperkingen van het toepassingsdomein van (symbolische) expertsystemen naar voren. Omdat expertsystemen niet op expertise-niveau beslissingen kunnen nemen of oordelen kunnen vormen, is het niet te rechtvaardigen om ze in te zetten voor taken die expertise vereisen. Dreyfus is echter wel overtuigd dat expertsystemen vaak een beperkt niveau van *competentie* kunnen bereiken: een prestatieniveau dat een beginnersniveau voorbijstreeft en vergelijkbaar is met dat van een gevorderde student. Wanneer prestaties op expert-niveau niet vereist zijn kunnen expertsystemen daarom wel hun nut hebben.

Een vraag die Dreyfus hier echter niet aansnijdt, is hoe beslist kan worden of een bepaalde taak expertise of slechts competentie vereist. Het bepalen van de juiste strafmaat voor een vergrijp vereist, nemen wij aan, de expertise van een rechter. Een rechter heeft de expertise om met inachtneming van de verschillende omstandigheden die golden bij het vergrijp en van de achtergrond van de verdachte, een redelijke strafmaat vast te stellen. Een wetgevende instantie kan echter ook beslissen dat rechters voortaan de strafmaat vaststellen op basis van een beperkt aantal formele principes, zoals de aard van het vergrijp, het strafblad van de vergrijper en een beperkt aantal andere, formeel toetsbare criteria. Op deze manier wordt het intuïtieve oordeel van de rechter uitgeschakeld en wordt zijn taak beperkt tot het toepassen van een aantal formele regels. Deze taak kan dan goed door een competent expertstelsel worden overgenomen.

Of het gebruik van expertsystemen in bepaalde toepassingsdomeinen te rechtvaardigen is, hangt dus mede af van een keuze om deze domeinen te formaliseren en het intuïtieve oordeel er van uit te sluiten. De AI-onderzoeker en -criticus Weizenbaum schreef in 1976 reeds een invloedrijke 'kritiek van de instrumentele rede,' waarin hij de neiging bekritiseert om menselijke problemen te reduceren tot berekenbare, logische problemen. Ook zonder de computer treedt dit verschijnsel al op, maar de inzetbaarheid van computers verschaft nog een extra excuus voor dergelijke pogingen tot formalisering. Weizenbaums conclusie, die Dreyfus stellig zal onderschrijven, is dat ook in specialistische domeinen het intuïtieve oordeel van mensen onmisbaar is.⁴

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

afhankelijk van de strategie van de tegenstander, en deze is aan verandering onderhevig. Schakers kunnen schaakcomputers daarom 'verrassen' door nieuwe strategieën te ontwikkelen waarop het computerprogramma onvoldoende is afgestemd.

⁴ Een meer recente en uitvoeriger kritiek van expertsystemen wordt geboden in Collins (1990). Zie voor ethische discussies over expertsystemen Forester & Morrison (1994, hoofdstuk 7) en Van den Hoven (1995, hoofdstuk 4).

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 vergaand type ITS neemt een actief begeleidend 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. Hiervoor moet een ITS beschikken over een bepaalde hoeveelheid didactische vaardigheden.

Een 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. Volgens Dreyfus zijn ITS-en daarom 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 halfgevorderden 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 halfgevorderden 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.⁵

7. Conclusie: Het gelijk en de invloed van Dreyfus

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 kunnen interpreteren, een robot over obstakels laten heenklimmen, of creativiteit vereisende problemen kunnen oplossen. Voor Dreyfus' kritiek van de neurale netwerken-AI is het nog te vroeg om te zeggen of deze hout snijdt, maar tot nu toe is het door Dreyfus omschreven generalisatieprobleem daar nog onopgelost.

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. Het geldt ook voor het invloedrijke werk van Agre en Chapman aan MIT (Agre 1988; Chapman 1991), soms genoemd 'Heideggeriaanse AI' vanwege het feit dat het een aantal door Heidegger en Dreyfus gepropageerde gezichtspunten probeert te implementeren in het AI-onderzoek, zoals het feit dat intelligentie gesitueerd is in een wereld en geen regels vereist en dat handelingen doelgericht kunnen zijn zonder dat er expliciete doelen hoeven te zijn gerepresenteerd.

De gesitueerdheid van intelligentie is ook 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 stoelen op Heideggeriaanse uitgangspunten.⁶ 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. Uitgangspunt in dit onderzoeksproject is dat menselijke intelligentie menselijke interacties met de wereld vereist en daarom een lichaam waarmee zulke interacties mogelijk zijn (Brooks & Stein 1994). Cog is een robot die is uitgerust met kunstmatige zintuigen (inclusief sensoren om de positie van het eigen lichaam te bepalen), een stem, en aanstuurbare ledematen. De 'hersenen' van Cog bestaan uit een gedeeltelijk parallel computersysteem. Het is de bedoeling dat Cog door zijn sensomotorische interacties met de omgeving sensomotor-intelligentie verwerft, om op basis hiervan 'hogere' vormen van intelligentie te ontwikkelen.⁷

⁵ Het door Dreyfus gesignaleerde gevaar dat mensen zichzelf gaan zien als computers is al bewaarheid, zoals blijkt uit psychologisch onderzoek van Sherry Turkle (1984). Voor een filosofische discussie over de implicaties van informatietechnologie voor het zelfbeeld van de mens, zie Coolen (1992).

⁶ Zie o.a. Winograd & Flores (1986) en Winograd (1995); Zie ook het invloedrijke werk van Suchman (1987).

⁷ Ook in de psychologie en cognitiewetenschap heeft een conceptie van intelligentie als een gesitueerd en belichaamd gegeven de laatste tien jaar veld gewonnen (b.v. Clark 1996; Varela, Thompson & Rosch 1991; Johnson 1987; Lakoff 1987)

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.

BIBLIOGRAFIE

Agre, Philip

1988, *The Dynamic Structure of Everyday Life*, MIT AI Lab Technical Report 1085.

Anderson, James

1995, *An Introduction to Neural Networks*, MIT Press, Cambridge, MA.

Brooks, Rodney, en Stein, Linda

1994, *Building Brains for Bodies*, *Autonomous Robotics* 1, 7-25.

Chapman, David

1991, *Vision, Instruction, and Action*, MIT Press, Cambridge, MA.

Clark, Andy

1996, *Being There. Putting Brain, Body, and World Together Again*, MIT Press, Cambridge, MA.

Collins, H. M.

1991, *Artificial Experts*, MIT Press, Cambridge, MA.

Coolen, T. M. T.

1992, *De machine voorbij. Over het zelfbegrip van de mens in het tijdperk van de informatietechniek*, Boom, Meppel en Amsterdam.

Crevier, Daniel

1993, *AI: The Tumultuous History of the Search for Artificial Intelligence*, Basic Books, New York.

Dreyfus, Hubert L.

1965, *Alchemy and Artificial Intelligence*, The RAND Corporation Paper P-3244.

1967, *Why computers must have bodies in order to be intelligent*, *Review of Metaphysics* 21, 13-32.

1972, *What Computers Can't Do: A Critique of Artificial Reason*, Harper and Row, New York.

1991, *Being-in-the-World: A Commentary on Heidegger's Being and Time*, MIT Press, Cambridge, MA.

1992, *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, Cambridge, Massachusetts.

1996, *Response to my critics*, *Artificial Intelligence* 80, 171-191.

Dreyfus, Hubert L. en Dreyfus, Stuart E.

1986, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the*

- Computer*. New York, Free Press.
- 1988, Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint, *Daedalus* 117, 15-43.
- Forester, Tom en Morrison, Perry
1994, *Computer Ethics: Cautionary Tales and Ethical Dilemma's in Computing*, MIT Press, Cambridge, MA.
- Hoven, M. J. van den
1995, *Information Technology and Moral Philosophy*. Proefschrift Erasmus Universiteit, Rotterdam.
- Johnson, Mark
1987, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL, University of Chicago Press.
- Lakoff, George
1987, *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*, Chicago, IL, University of Chicago Press.
- Meijering, Theo
1993, Neuraal vernuft en gedachteloze kennis. Het moderne pleidooi voor een niet-propositioneel kennismodel, *Algemeen Nederlands tijdschrift voor wijsbegeerte* 85, 24-48.
- Suchman, Lucy
1987, *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, Cambridge University Press.
- Turkle, Sherry
1984, *The second self: computers and the human spirit*, New York, Simon & Schuster. [NL Het tweede ik: computers en de menselijke geest. Groningen: Wolters-Noordhoff, 1986]
- Varela, Francisco, Thompson, Evan, en Rosch, Eleanor
1991, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge, MA.
- Weizenbaum, Joseph
1976, Computerkracht & Mensenmacht. Van Oordeel tot Berekening. (1984) Amsterdam, Contact. (*Computer Power and Human Reason*, Freeman, San Francisco.)
- Winograd, Terry
1995, Heidegger and the Design of Computer Systems, in: Feenberg, Andrew en Hannay, Alastair, *Technology and the Politics of Knowledge*, Bloomington and Indianapolis, Indiana University Press.
- Winograd, Terry, en Flores, Fernando
1986, *Understanding Computers and Cognition: A New Foundation for Design*, Ablex, Norwood, NJ.