COLIN BROWN, PETER HAGOORT EN THEO MEIJERING (RED.)

Vensters op de geest

Cognitie op het snijvlak van filosofie en psychologie

De Grafietreeks staat onder redactie van Aad Blok, Colin Brown, Radboud Engbersen, Peter Hagoort, Sjaak Koenis, Theo Meijering, Janneke Plantenga, Lies Wesseling, Pauline Westerman

JAN SLEUTELS EN BART GEURTS

Knopen en connecties

Filosofische aspecten van het connectionisme

Sinds enkele jaren bestaat er in de cognitiewetenschap een zogenaamde 'connectionistische' school. Het connectionisme onderscheidt zich op tal van punten van het traditionele, computationalistische denken; het is daarom niet zonder belang de verhouding tussen connectionisme en computationalisme nader te onderzoeken. Zijn zij met elkaar in tegenspraak? Zo ja, welk van beide heeft dan de beste kaarten? Zo neen, hoe moet hun relatie dàn worden gezien?

Na een korte schets van de filosofisch gezien meest belangwekkende eigenschappen van connectionistische modellen, zullen wij in dit artikel het connectionisme en het computationalisme op kernpunten met elkaar vergelijken. Vervolgens bespreken wij een drietal onderscheiden visies op de verhouding tussen connectionisme en computationalisme, die als respectievelijk de 'subsymbolische', de 'implementationele' en de 'eliminatieve' interpretatie van het connectionisme kunnen worden aangemerkt. Wij trachten aan te tonen dat elk van deze posities gebreken kent en argumenteren voor een alternatieve, 'revisionistische' kijk op het connectionisme.

Knopen en connecties

Connectionistische modellen bestaan uit netwerken van onderling verbonden elementaire verwerkingseenheden die hier verder kortweg *knopen* zullen worden genoemd. Langs de verbindingen tussen de knopen worden signalen (prikkels) doorgegeven. Welk effect een prikkel heeft op de ontvangende knoop is onder meer afhankelijk van de aard van de verbinding. De verbindingen tussen de knopen hebben elk een bepaald gewicht. Prikkels die langs negatieve (inhiberende) verbindingen worden aangevoerd zullen het activatieniveau van de ontvangende knoop verlagen; hoe negatiever de verbinding, des te lager het activatieniveau. Positieve verbindingen verhogen het activatieniveau van de aangestuurde knoop; hoe positiever de verbinding, des te hoger het activatieniveau. Overschrijdt de activatie van een knoop een bepaalde drempelwaarde, dan geeft de knoop een prikkel door naar elk van de andere knopen waarmee hij is verbonden.

Op elk gegeven ogenblik kan de toestand waarin een netwerk zich bevindt worden weergegeven als een configuratie van activatieniveaus van de individuele knopen. Zodra het netwerk een input krijgt toegevoerd, d.w.z. zodra een of meer (zogeheten input-)knopen worden geactiveerd, spreidt de activatie zich over het netwerk uit Hoe de toestand van het netwerk zich van moment tot moment wijzigt (de *dynamiek* van het netwerk) wordt bepaald door de structuur van het netwerk, d.w.z. door de plaats en het gewicht van de verbindingen en de drempelwaarde van de knopen. De dynamiek van het netwerk kan mathematisch worden weergegeven in de vorm van differentiaalvergelijkingen waarin de toestandswijziging van het systeem een functie van de tijd is. Gegeven een bepaalde configuratie van activiteit op tijdstip t, kan uit de dynamische structuur van het netwerk worden berekend hoe de configuratie van activiteit op tijdstip 1 = t + l er uit zal zien. De output van het netwerk wordt verzorgd door zogenaamde outputknopen. Deze zijn zodanig verbonden met de rest van het netwerk dat zij selectief gevoelig zijn voor (d.w.z. hun drempelwaarde alleen bereiken bij) bepaalde configuraties van activiteit in het netwerk; doorgaans zijn dit de min of meer stabiele toestanden van activiteitsverdeling waarin het netwerk na verloop van tijd 'tot rust komt'. Volledigheidshalve merken wij op dat niet alle connectionistische modellen in het bezit zijn van drempelwaarden. Wij gaan hier echter uit van de meest abstracte vorm van een netwerk waarin alle knopen drempelwaarden bezitten en met elkaar zijn verbonden.

De conceptuele **vectorruimte** van het **connectionisme**

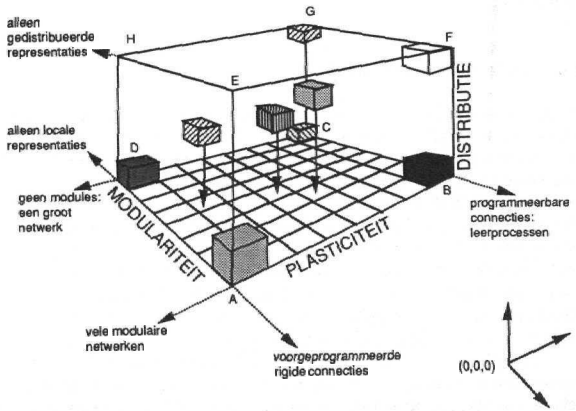
Bovenstaande ruwe schets van het connectionisme moge gelden als de 'harde kern" van de nieuwe onderzoeksrichting.[1] Aangezien het ons hier te doen is om bepaalde filosofisch interessante algemene eigenschappen van connectionistische modellen, is het van belang dat het geschetste theoretische kader in elk geval ruim genoeg is om het merendeel van de vele uiteenlopende soorten van connectionistische netwerken te omvatten. Binnen dit kader bestaat een grote mate van vrijheid in de keuze van meer specifieke theoretische invullingen van het boven geschetste basisidee. Al naar gelang de specifieke eigenschappen waarmee de netwerken in de diverse concrete uitwerkingen worden bedeeld, kunnen machines ontstaan die evenveel op elkaar lijken als een loopfiets en een Maserati. Enkele der belangrijkste parameters bij de implementatie van het connectionistische basisidee zijn ongetwijfeld de *modulariteit*, de *plasticiteit* en de *distributie van representatie* van de systemen. Onder *'modulariteit'* verstaan wij de mate waarin een netwerk is opgebouwd uit zelfstandig werkende deelnetwerken. Onder 'plasticiteit' verstaan wij de mate waarin een netwerk in staat is

tot leren, door zijn structuur (de gewichten van de verbindingen en de drempelwaarden van de knopen) te wijzigen. Op de modulariteit en plasticiteit van connectionistische en andere cognitieve modellen wordt uitvoerig ingegaan in andere bijdragen in deze bundel.[2] Wij zullen ons hier in het bijzonder wijden aan het probleem van de *semantiek* van connectionistische modellen, d.w.z. aan de kwestie van (locale en gedistribueerde) *representaties*, en dan met name aan de vraag of connectionistische modellen in dit opzicht wezenlijk verschillen van traditionele computationalistische modellen van cognitie.

Om de lezer een globale indruk te geven van de ideeenwereld van het connectionisme kunnen wij de genoemde drie parameters voorstellen als onafhankelijke assen in een geometrische ruimte; de vrijheid van theoretiseren rond de harde kern van het connectionisme kan dan worden weergegeven als een driedimensionale 'conceptuele vectorruimte' van connectionistische modellen, zoals afgebeeld in figuur 1. Specifieke modellen, gekenmerkt door een bepaalde mate van modulariteit, een bepaalde mate van plasticiteit en een bepaalde mate van distributie van representaties, worden weergegeven door een vector in deze ruimte. Minder specifieke modellen, waarin de keuze voor een bepaalde invulling van de parameters nog in meerdere of mindere mate wordt opengelaten, worden weergegeven door lichamen in de conceptuele ruimte.[3]

Connecüonistische modellen stellen zich ten doel netwerken te beschrijven die een (meer of minder precies omschreven) cognitieve functie kunnen vervullen. Dit betekent onder meer dat de activiteit van een netwerk begrepen moet kunnen worden in termen van psychologische, cognitief relevante generalisaties. Deze generalisaties zijn doorgaans van semantische aard, d.w.z. zij brengen de activiteit van het netwerk in verband met de interactie van het systeem met zijn omgeving. De activiteit van het netwerk moet derhalve *semantisch interpreteerbaar* zijn. Toestanden van het netwerk en/of van zijn delen moeten kunnen worden opgevat als *representaties* van objecten en eigenschappen in de *'buitenwereld'*, d.w.z. in het deel van de werkelijkheid waarop het door het netwerk belichaamde kennisdomein wordt geacht betrekking te hebben.

Zoals afgebeeld in figuur l, bestaan er twee algemene strategieën om de activiteit van een netwerk semantisch te interpreteren. De zogeheten *localistische* strategie (figuur l, vlak ABCD) kent interpretaties toe aan afzonderlijke knopen in het netwerk. Hier kan onder meer worden gedacht aan Rumelhart en McClelland's (elders in deze bundel beschreven) model voor woord- en letterherkenning, waarin de afzonderlijke knopen in de drie lagen respectievelijk staan voor onderdelen van letters (lijnsegmenten), letters (A, B, etc.), en woorden (ADEL, AMBT, etc.).[4] De interpretatie van localistische netwerken als geheel is een functie van die van de toestand van de afzonderlijke knopen.



Figuur 1: Een deel van de conceptuele vectorruimte van het connectionisme.

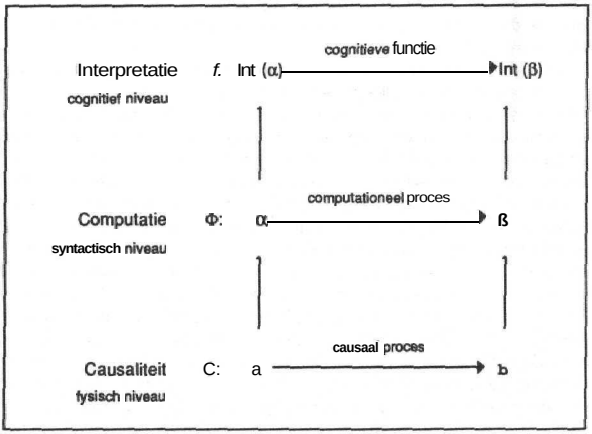
fa de huidige connectionistische theorievorming is het vooral een tweede, zogenaamde *gedistribueerde* interpretatiestrategie die veel aandacht krijgt. In gedistribueerde netwerken (gelegen in vlak EFGH, figuur 1) zijn niet zozeer de afzonderlijke knopen eenheden van interpretatie, als wel clusters van (verspreide) knopen. Eigenschappen en objecten worden gerepresenteerd als geaggregeerde activatiepatronen van meerdere over het netwerk verdeelde knopen. Daargelaten of de activiteit van afzonderlijke knopen überhaupt zinvol kan worden geïnterpreteerd, is het de verdeling van activiteit over clusters van knopen die wordt geïnterpreteerd als een representatie van objecten of eigenschappen.

De orthodoxe computationalist

In de klassieke traditie op het gebied van de computationele theorie van cognitie worden kennisverschijnselen verklaard op grond van de hypothese dat de cognitieve functies (vermogens, gedragsrepertoires, enzovoorts) van een systeem worden gerealiseerd door computationele processen die zich afspelen in de fysische structuur van het systeem. De basisgedachte is heel simpel. De fysische structuur van een cognitief systeem kan systematisch worden geïnterpreteerd als een systeem van symbolen. Bepaalde processen in de structuur grijpen aan op deze symbolen en produceren nieuwe symbolen. De ingaande en uitgaande symbolen kunnen systematisch worden geïnterpreteerd als de input en output van de te verklaren cognitieve functie.

In figuur 2 is deze computationalistische verklaringsstrategie schematisch weergegeven. Op *fysisch* niveau vinden causale processen plaats waarbij een toestand *a* systematisch wordt omgezet in een toestand *b*. In een computer, bijvoorbeeld, spelen zich op fysisch niveau electrodynamische processen af waardoor de verdeling van electrische lading over myriaden microscopische onderdelen van siliciumchips wordt gewijzigd. Analoog hieraan spelen zich in de hersenen van de mens electrochemische processen af waardoor de activatietoestand van ontelbare zenuwcellen zich voortdurend wijzigt.

Deze causale processen kunnen relatief abstract worden weergegeven als een functie Φ *die* ß (een abstractie over bepaalde eigenschappen van de eindtoestand b) berekent uit a (een abstractie over bepaalde eigenschappen van de aanvangstoestand a). Dit beschrijvingsniveau zouden wij syntactisch kunnen noemen. Het onderscheidt zich van het vorige niveau doordat de begin- en eindtoestand niet worden beschreven als oorzaak en gevolg, maar als argument en functiewaarde die volgens een bepaalde calculus van rekenregels met elkaar samenhangen. Anderzijds onderscheidt het zich van het hierna te noemen semantische niveau doordat de berekeningen louter formele symbolen



Figuur 2: Computatie en interpretatie.

manipuleren die (nog) geen cognitieve betekenis hebben. Een syntactische beschrijving van de processen die zich in een computer afspelen wordt gegeven in 'machinetaal'; deze geeft de toestand van de computer weer als een verdeling van enen en nullen over diverse 'registers', en beschrijft hoe de ene verdeling wordt berekend uit de andere. Analoog hieraan kunnen volgens het computationalisme ook processen in het menselijk zenuwstelsel worden beschreven in een 'machinetaal' van de geest; zelfs een individuele zenuwcel kan abstract worden beschreven als een calculator die de netto som van positieve en negatieve prikkels berekent, deze vervolgens vergelijkt met een bepaalde drempelwaarde, en bij een voldoende resultaat zelf ook een prikkel afgeeft naar zijn buurcellen.

Op *cognitief* niveau, tenslotte, kunnen deze syntactische processen worden geïnterpreteerd als de berekening van een cognitieve functie Ψ van de *interpretatie* van α naar de *interpretatie* van ß. Onder deze interpretatie krijgen a en ß een representationele *inhoud.* De computationele processen kunnen derhalve worden beschreven als operaties over *symbolen,* d.w.z. *betekenisdragers.* In een computer kan doorgaans het in bijvoorbeeld Lisp, Prolog of Pascal geschreven programma worden aangewezen als het laagste niveau van 'cognitieve' interpretatie. Het programma beschrijft wat er eigenlijk gebeurt wanneer, onzichtbaar voor de gebruiker, berekeningen op reeksen enen en nullen worden uitgevoerd, bijvoorbeeld dat de zin 'Kuifje is vindingrijk' wordt ontleed in het onderwerp *Kuifje* en het gezegde is *vindingrijk.* De zuiver formele operaties in machinetaal krijgen aldus een betekenis in termen van het gebruik dat de man of vrouw achter het toetsenbord ervan kan maken. Volgens het computationalisme geldt voor de menselijke hersenen in wezen hetzelfde. De berekeningen die, onzichtbaar voor *de* 'gebruiker', door het zenuwstelsel op syntactisch niveau worden gemaakt, kunnen op het niveau van ons bewustzijn stelselmatig worden geïnterpreteerd in termen van cognitieve processen als waarnemen, redeneren, zich herinneren en dergelijke meer.

Laten wij een simpel voorbeeld nemen: wil een systeem kunnen worden beschreven als een inferentiemechanisme, dan zal de werking ervan moeten kunnen worden beschreven als het berekenen van symbolen die kunnen worden geïnterpreteerd als *conclusies* uit symbolen die kunnen worden geïnterpreteerd als *premissen.* Aan dit voorbeeld kan nog een volgende les worden verbonden. De relatie tussen premissen en conclusie in een geldige redenering Iaat zich nader analyseren in termen van de inwendige structuur van de onderdelen ervan. Dat uit (AAC en $A \rightarrow B$) volgt dat B, kan nader worden begrepen in termen van de elementen 'A', 'A', '$\rightarrow$', 'B' en 'C'. (In de hier gebruikte standaardnotatie van logische symbolen moet '$A \wedge C$' worden gelezen als 'A en C', '$A \rightarrow B$' als 'indien A dan B'.) De symbolen hebben een *combinatorische structuur:* met behulp van een aantal eenvoudige formatieregels kunnen uit atomaire symbolen zoals 'A', 'A' en 'C' complexe symbolen

zoals '$A \wedge C$' worden gevormd; de eigenschappen van complexe symbolen zijn daarbij een functie van de eigenschappen van de samenstellende atomaire symbolen. Welnu, volgens het klassieke computationalisme corresponderen met deze structurele relaties op semantisch niveau (het niveau van de *interpretatie* van fysische systemen) soortgelijke structurele relaties op syntactisch niveau (het niveau van de *computationele* processen) en op fysisch niveau (het niveau van de causale processen). Wanneer in het zojuist gegeven voorbeeld $Int(\alpha) = (A \wedge C, A \rightarrow B)$, dan is volgens het computationalisme de fysische structuur α samengesteld uit deelstructuren als 'A', V, 'C', enzovoorts, die ook elk afzonderlijk kunnen worden geïnterpreteerd. In strikte zin zijn het deze atomaire deelstructuren waarop computationele processen aangrijpen; hoe een complex symbool *computationeel wordt* verwerkt is een functie van de verwerking van de samenstellende delen.

Ter onderscheiding van de semantische en de niet-semantische eigenschappen van fysische symbolen, d.w.z. ter onderscheiding van hun *interpretatie* en hun *computatie*, is het gebruikelijk de computationele structuur van fysische symbolen aan te duiden als hun *syntaxis*. Het computationalisme kan zo in een notedop worden samengevat als de stelling dat de *semantiek* van cognitieve processen wordt weerspiegeld in de *syntaxis* van de onderliggende fysische processen. In *de* woorden van Fodor en Pylyshyn:

> If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a syntactically driven machine whose state transitions satisfy semantic criteria of coherence. Such a machine would be just what's required for a mechanical model of the semantical coherence of thought; correspondingly, the idea that the brain is such a machine is the foundational hypothesis of Classical cognitive science.[5]

Vanwege hun syntactische structuur vertonen computationele systemen zoals hersenen en computers grote overeenkomst met natuurlijke en formele talen. Hun grammatica van formatie- en transformatieregels voor de verwerking van atomaire en complexe symbolen suggereert dat mentale processen een subtiele vorm van inwendig spreken zijn. Het computationalisme kan dan ook kortweg gedefinieerd worden als de veronderstelling dat er een *language of thought* bestaat.[6]

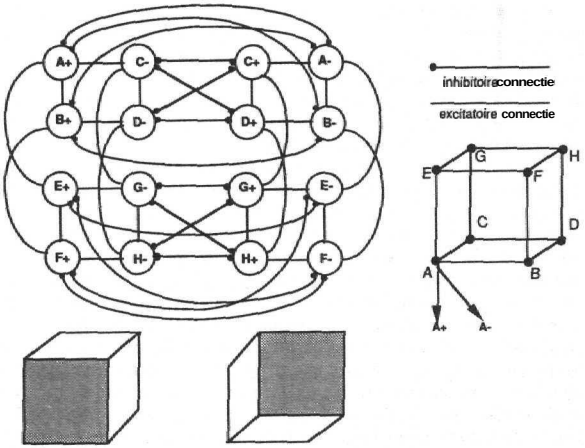
Connectionistisch computationalisme?

Wanneer wij nu de eerder gegeven profielschets van het connectionisme vergelijken met die van het orthodox computationalisme, lijkt het

connectionisme op het eerste gezicht een loepzuivere vorm van computationalisme te zijn. Betrekkelijk modulair geordende netwerken, zo hebben wij verondersteld, berekenen betrekkelijk specifieke cognitieve functies. Deze cognitieve functies worden verklaard in termen van onderliggende computationele processen op fysische symbolen. Aan het netwerk wordt een input a toegevoerd (een bepaalde configuratie van activatie van de knopen), door de knopen en connecties wordt een bepaalde functie Φ van a naar ß berekend (mathematisch beschrijfbaar in de vorm van differentiaalvergelijkingen), en het resultaat is een output ß (een nieuwe verdeling van activatie over de knopen). Op semantisch niveau kan de input/outputfunctie zodanig worden geïnterpreteerd dat het systeem *de* te verklaren cognitieve functie Ψ van $\mathrm{Int}(\alpha)$ naar $\mathrm{Int}(\beta)$ berekent. Computationalisme *pur sang'*.

Wanneer wij hier beweren dat het connectionisme zich, althans op het eerste gezicht, voordoet als een vorm van orthodox computationalisme, bedoelen wij daarmee zeker niet dat het ook in alle opzichten overeenkomt met de symbolistische modellen uit de traditionele cognitieve psychologie en AI. Er bestaan immers overduidelijke verschillen tussen beide soorten van modellen. Doorgaans worden deze verschillen toegespitst op het bestaan van expliciet, locaal gerepresenteerde symbolen, en regels voor de verwerking van deze symbolen in traditionele cognitieve modellen, en het ontbreken van dergelijke expliciete regels en representaties in connectionistische modellen? Het ontbreken van expliciet gerepresenteerde regels voor de berekening van symbolen betekent echter nog niet noodzakelijk dat er *in het geheel geen* berekeningen op symbolen plaatsvinden.

Laten wij aan de hand van een concreet voorbeeld nagaan in welke zin connectionistische modellen een vorm van Computationalisme kunnen worden genoemd. In figuur 3 is een simpel netwerk voor de interpretatie van de bekende kubus van Necker weergegeven.[8] Elke knoop in het netwerk kan worden geïnterpreteerd als een bepaalde hypothese over de ruimtelijke stand van een van de hoekpunten van de kubus. Het label 'H+' op een knoop staat voor de hypothese dat hoekpunt H op de voorgrond ligt, d.w.z. naar de waarnemer toe wijst; het label 'H-' staat voor de hypothese dat hoekpunt H op de achtergrond ligt, d.w.z. van de waarnemer af gericht is. De verbindingen tussen de knopen weerspiegelen de diverse semantische relaties tussen de hypotheses waarvoor de knopen staan. Wanneer bij aanwezigheid van een bepaalde eigenschap M (bijvoorbeeld dat hoekpunt A naar de waarnemer wijst) normaliter mag worden verwacht dat ook eigenschap *N* aanwezig zal zijn (bijvoorbeeld dat ook hoekpunt B naar de waarnemer toe gericht is), moet er een positieve verbinding bestaan tussen de knoop die de hypothese dat M aanwezig is belichaamt en de knoop die staat voor de hypothese dat *N* aanwezig is. Omgekeerd, wanneer in de normale omgeving van het systeem geldt dat *N* niet



Figuur 3: Een eenvoudig netwerk voor de waarneming ran de kubus van Necker.

aanwezig zal zijn indien M aanwezig is (bijvoorbeeld dat hoekpunt B niet naar achteren gericht zal zijn indien hoekpunt A naar voren gericht is), dient de verbinding tussen de desbetreffende knopen inhiberend te zijn. Zijn de restricties statistisch gezien klein, dan zal de verbinding zwak moeten zijn; zijn de restricties groot, dan zal de verbinding navenant sterk moeten zijn.

Naast deze 'structurele' informatie, berustend op een a priori geometrische analyse van de mogelijke stand van kubussen, rechtstreeks gerepresenteerd in de verbindingen tussen de knopen, kunnen ook de op een bepaald moment aan het netwerk toegevoerde inputwaarden worden geïnterpreteerd als symbolen. Een positieve input naar een knoop betekent dat er reden is om te veronderstellen dat de door de knoop gerepresenteerde hypothese over de buitenwereld correct is; hoe groter de inputwaarde, des te sterker de hypothese. Omgekeerd betekent een negatieve input naar een knoop dat er reden is om te veronderstellen dat de desbetreffende hypothese onjuist is; hoe lager de waarde, des te zwakker de hypothese.

Deze twee soorten van representaties, de 'structurele' connectiewaarden en de 'incidentele' inputwaarden, bepalen hoe de verdeling van activiteit over het netwerk zich verder zal ontwikkelen. Het netwerk maakt een balans op van de gerepresenteerde hypotheses. Het zal het aantal actieve, tegenstrijdige hypotheses proberen te minimaliseren en het aantal actieve coherente en elkaar versterkende hypotheses proberen te maximaliseren. Uiteindelijk komt het netwerk, via een procedure die bekend staat als 'relaxatie', terecht in een stabiele toestand waarin aan zoveel mogelijk restricties wordt voldaan, met dien verstande dat sterke restricties hierbij prioriteit krijgen boven zwakke restricties. Het netwerk van figuur 3 kent twee stabiele toestanden, de configuraties (A+, C-, B+, D-, E+, G-, F+, H-) en (A-, C+, B-, D+, E-, G+, F-, H+), corresponderend met de twee consistente interpretatiemogelijkheden van de kubus.'

De les uit dit voorbeeld zal duidelijk zijn: net als orthodoxe computationele systemen kan het netwerk van figuur 3 worden beschouwd als een systeem van fysische symbolen, in dit geval knopen en connecties die de geometrie van een kubus en zijn delen representeren. Er bestaat een expliciet verband tussen processen op syntactisch niveau en processen op semantisch, cognitief niveau. Wanneer het systeem bijvoorbeeld wordt gevoed met de premisse dat hoekpunt C in het voorvlak ligt, trekt het de conclusie dat de kubus zich in de rechtsonder afgebeelde stand bevindt. Hoe dit mogelijk is, wordt verklaard als een computationeel proces: *de* activatie van knoop C+ verspreidt zich over het netwerk tot de stabiele toestand {A-, C+, B-, D+, E-, G+, F-, H+} is bereikt.

Connectionisme onder de microscoop

Wanneer wij de vergelijking tussen connectionisme en orthodox computationalisme vrij abstract houden, zoals tot dusver is gedaan, bestaart er duidelijke overeenkomsten tussen beide: in grote lijnen lijkt connectionisme een vorm van computationalisme te zijn. Kijken wij daarentegen iets preciezer, dan dringt zich een aantal vragen op. Zijn de configuraties fA+, C-, B+, D-, E+, G-, F+, H-) en (A-, C+, B-, D+, E-, G+, F-, H+} *gedistribueerde* representaties van de respectievelijk links- en rechtsonder afgebeelde kubussen, of zijn het verzamelingen *locale* representaties van hoekpunten met een bepaalde oriëntatie? Is (A+, C-, B+, D-, E+, G-, F+, H-, A-, C+, B-, D+, E-, G+, F-, H+) een representatie van een kubus met indifferente oriëntatie? Is (A+) een locale representatie van een hoekpunt dat naar de waarnemer toe wijst, of is het slechts een onderdeel van een grotere gedistribueerde representatie? Is |A+, C-) een locale of een gedistribueerde representatie van hoekpunt A dat naar voren wijst, of een locale of gedistribueerde representatie van ribbe AC? Is {A+, A-) een representatie van een hoekpunt A met indifferente positie? Het lijkt onmogelijk die vragen precies te beantwoorden zolang niet exact gedefinieerd is wat onder 'locaal' en 'gedistribueerd' moet worden verstaan. Misschien moeten wij zeggen dat locaal' en 'gedistribueerd' relatieve begrippen zijn: misschien maakt het niet uit hoe wij een representatie noemen, als wij er maar bij zeggen *ten opzichte waarvan* zij locaal of gedistribueerd is. Immers, zolang de afzonderlijke knopen die zijn betrokken bij een zogenaamde gedistribueerde representatie geïnterpreteerd kunnen worden in termen van hun relatie tot onderdelen of aspecten ('microfeatures') van de gerepresenteerde objecten (of eigenschappen, feiten, enzovoorts), is elke gedistribueerde representatie een geordende verzameling of 'vector' van locale representaties. Volgens deze redenering hangt het enkel van het oplossend vermogen van de analyse af of een representatie locaal dan wel gedistribueerd moet worden genoemd.

Een andere, filosofisch interessantere mogelijkheid is dat de vraag of representaties locaal dan wel gedistribueerd zijn inderdaad afhangt van de eenheden van interpretatie, maar dat er een meest *natuurlijke* eenheid van interpretatie bestaat, een eenheid die het meest geschikt is voor de verklaring van cognitieve processen. Dit is in wezen de positie van de orthodoxe, symbolistische computationalist. Het idee is daarbij dat cognitieve machines (of het nu netwerken, hersenen of von Neumann-computers zijn) weliswaar op tal van niveaus van functionele analyse kunnen worden beschreven, maar dat er één homogeen, meest elementair niveau van zinvolle interpretatie is: wat onder dit niveau ligt is niet-symbolische hardware, wat erboven ligt is software samengesteld uit de atomaire symbolen. Het niveau van *natuurlijke symbolen,* zoals wij ze zouden kunnen noemen, bepaalt waar de relevante com-

putationele processen zich afspelen. Weliswaar vinden ook berekeningen plaats *onder* dit niveau, in de niet-symbolische of (zoals wij hier voortaan zullen zeggen:) 'sub-symbolische' hardware, maar dat zijn geen berekeningen van *symbolen*. Voor de berekeningen die plaatsgrijpen *boven* het niveau van natuurlijke symbolen geldt dat zij een functie zijn van de onderliggende natuurlijke computaties op atomaire symbolen.

Gesteld dàt er zoiets als 'natuurlijke symbolen' bestaan, dan hoeft dat op zich nog niet te betekenen dat er *één, homogeen* niveau van natuurlijke symbolen is, identiek voor alle cognitieve processen. Misschien ligt de grens tussen symbolische en subsymbolische processen voor verschillende cognitieve functies op verschillende niveaus van abstractie. Zo is het alleszins voorstelbaar dat het menselijk brein van origine een connectionistisch netwerk is dat, in een betrekkelijk laat stadium van zijn evolutie, heeft 'ontdekt' dat het voor sommige doeleinden evolutionair gesproken aantrekkelijker is om een von Neumann-machine te simuleren.[10] Ons bewuste redeneervermogen, bijvoorbeeld, zou in dit opzicht wellicht beter kunnen worden begrepen in termen van zijn virtuele von Neumann-architectuur dan in termen van de connectionistische hardware waarop de virtuele machine draait. Andere, primitievere cognitieve processen, daarentegen, zouden wellicht juist beter kunnen worden begrepen in termen van die connectionistische architectuur en de algoritmen die zich op netwerkniveau afspelen. Afhankelijk van het domein van cognitie dat wordt onderzocht, zou dan van geval tot geval moeten worden nagegaan of de desbetreffende processen meer als 'hardware' dan wel meer als 'software' moeten worden geanalyseerd, d.w.z. als processen op netwerk-niveau dan wel als processen op een hoger niveau van analyse.

Op de vraag of er een *uniform* niveau van natuurlijke symbolen bestaat zullen wij hier niet verder ingaan. Wij richten ons op de fundamentelere vraag naar de relatie tussen connectionistische netwerken en het orthodox computationalisme, uitgaande van het idee dat er een meest natuurlijk niveau van symbolen is, of dat nu voor alle kennisdomeinen hetzelfde is of niet. De vraag is dus niet waar de grens tussen symbolische en subsymbolische processen ligt, maar hoe connectionisme en computationalisme zich verhouden *indien er ergens* zo'n grens ligt. Welnu, uitgaande van het idee van een meest natuurlijk niveau van symbolen, hoeft het connectionisme niet noodzakelijk als een vorm van computationalisme te worden geïnterpreteerd: immers, de berekeningen die in netwerkmodellen worden weergegeven spelen zich wellicht af op een niveau dat boven of onder dat van de natuurlijke symbolen ligt. Wij kunnen hier een drietal mogelijkheden onderscheiden:

1. Wat berekend wordt zijn geen symbolen.
2. Wat gesymboliseerd wordt is niet berekend.
3. De berekende symbolen zijn geen 'cognitieve functie'.

Deze drie mogelijkheden zullen wij hier verder aanduiden als respectievelijk *subsymbolisme*, *implementationisme*, en *eliminisme*. Alvorens elk van deze drie posities in detail te bespreken, zullen wij ze eerst in het kort de revue laten passeren.

Het centrale idee van het subsymbolisme is dat mentale berekeningen niet worden uitgevoerd op symbolen, maar op 'kleinere' eenheden, zogenaamde 'subsymbolen'. Het subsymbolisme impliceert derhalve nog niet dat het symbolische niveau een hersenschim is; het symbolisch niveau wordt erkend als een soort macroscopisch bijverschijnsel van processen op subsymbolisch niveau. Een expliciete en nauwkeurige verklaring van cognitie kan evenwel alleen op het subsymbolisch niveau worden gegeven; een exacte wetenschap van symbolische representaties en processen zou bijgevolg onmogelijk zijn.

Terwijl het subsymbolisme kiest voor het subsymbolische niveau zonder daarmee het bestaan van symbolen te ontkennen, doet het implementationisme in feite het omgekeerde: het erkent het bestaan van subsymbolen, maar zoekt de verklaring van cognitieve verschijnselen juist in symbolen. Het implementationisme beschouwt het subsymbolische niveau als een realisering van het symbolische niveau, die vanuit het oogpunt van de cognitiewetenschap van ondergeschikt belang is. Zoals de hardware van een computer een programma implementeert, zo implementeren subsymbolen symbolen. En net zoals men over het algemeen het gedrag van een computer het beste kan begrijpen door de software (en niet de hardware) te bestuderen, zo zou men cognitie het beste kunnen verklaren aan de hand van symbolen (en niet van subsymbolen).

Zowel het subsymbolisme als het implementationisme erkennen, elk op zijn eigen wijze, het bestaan van zowel een symbolisch als een subsymbolisch niveau. Het eliminisme gaat er daarentegen van uit dat alleen het subsymbolisch niveau reëel is; het symbolisch niveau van mentale representaties is volgens deze voorstelling van zaken een hersenschim, die bijgevolg irrelevant moet zijn voor de verklaring van cognitieve verschijnselen.

Wij zullen nu nader ingaan op de diverse argumenten, problemen en mogelijkheden van de drie benaderingen van het connectionisme, zoals verdedigd door hun voornaamste vertegenwoordigers.

Wat **berekend** wordt zijn geen symbolen: **subsymbolisme**

Smolensky vergelijkt de verhouding tussen de berekeningen en de symbolen onder meer met die tussen de molecuulbewegingen en de

temperatuur van een gas.[11] Net zoals *de* temperatuur van een gas een zogenaamde *émergente* eigenschap is van de stochastische verdeling van de bewegingen van de afzonderlijke gasmoleculen, zouden ook de symbolen op conceptueel niveau een soort 'bijverschijnsel' zijn van de verdeling van activatie over groepen van knopen op computationeel niveau. 'Temperatuur' is weliswaar geen eigenschap van de afzonderlijke moleculen, maar wel van grote aantallen moleculen die in algemene, statistische termen worden beschreven; analoog hieraan zouden de afzonderlijke knopen en connecties in een connectionistisch netwerk geen representaties berekenen, maar zouden grote groepen knopen in algemene termen als representaties kunnen worden beschreven. "When connectionist computational systems are analyzed at higher levels, elements of symbolic computation appear as emergent properties."[12]

Aan de hand van de vergelijking tussen gas en netwerk kunnen wij twee onderscheiden aspecten van het subsymbolisme illustreren. Enerzijds beschrijven connectionistische modellen volgens Smolensky *hetzelfde* als symbolistische modellen, net zoals *de* fenomenologische en de statistische thermodynamica allebei de energietoestand van een gas beschrijven. Anderzijds beschrijven connectionistische modellen *preciezer* wat symbolistische modellen slechts *bij benadering* beschrijven, net zoals de statistische thermodynamica een preciezer beeld geeft van de energietoestand van de moleculen in een gas, die door de fenomenologische thermodynamica slechts *grosso modo* wordt beschreven. Enerzijds is er dus sprake van een *correspondentie* tussen elementaire eigenschappen en émergente eigenschappen, anderzijds is de relatie tussen beide slechts *approximatief*. Wij zullen ons hier eerst concentreren op het aspect van correspondentie; op het aspect van approximatie zullen wij later terugkomen.

Volgens Smolensky spelen de berekeningen in een connectionistisch model zich af *onder* het niveau van symbolen, en zijn de waarneembare symbolische eigenschappen in wezen betrekkelijk oppervlakkige bijverschijnselen. Hoe moet deze relatie tussen de émergente en de elementaire eigenschappen nu worden voorgesteld? In het geval van de temperatuur van een gas lijkt de verhouding nog betrekkelijk eenvoudig. Moleculen bewegen en botsen maar hebben geen temperatuur; het gas beweegt niet en botst niet maar heeft een bepaalde temperatuur. Er zijn echter ook andere soorten van systemen met émergente eigenschappen, waarin de relatie tussen 'onder-' en 'bovenbouw' veel complexer is. Een voorbeeld hiervan zijn orthodox-computationele systemen, die zich immers ook op deze wijze laten beschrijven. Op het niveau van de machinetaal kan de werking van een computer worden weergegeven in termen van berekeningen over activatietoestanden van registers (enen en nullen). Op hogere niveaus, bijvoorbeeld op dat van de symbolische algoritmes in Lisp of Prolog, vertonen groepen

van deze elementaire berekeningen émergente eigenschappen: zij kunnen worden beschouwd als computationele operaties op symbolen. Wanneer de elementaire eigenschappen van een emergent systeem berekeningen zijn, betekent dat dus nog niet automatisch dat de émergente eigenschappen *geen* berekeningen kunnen zijn. Wij kunnen Smolensky's vergelijking tussen gas en netwerk dan ook niet zomaar overnemen; als de symbolen inderdaad emergent zijn en de berekeningen inderdaad elementair zijn, volgt daaruit nog niet automatisch dat de symbolen niet worden *berekend*.

Laten wij de relatie tussen symbool en berekening nauwkeuriger bezien. Smolensky vat de huidige stand van zaken in het connectionistisch modelleren op dit punt als volgt samen:

> At present, each individual subsymbolic model adopts particular procedures for relating patterns of activity - activity vectors - to the conceptual-level descriptions of the inputs and outputs that define the model's task. The vectors chosen are often values of fine-grained features of the inputs and outputs, based on some pre-existing theoretical *analysis* of the *domain*.[13]

De relatie tussen de beschrijvingen op conceptueel niveau (de symbolen) en de beschrijvingen op subconceptueel niveau (de subsymbolen) moeten wij ons daarbij waarschijnlijk ongeveer voorstellen als de relatie tussen de beschrijving van een van de twee standen van de Neckerkubus en de beschrijving van de oriëntatie van de diverse hoekpunten van de kubus in figuur 3. De subsymbolen staan voor de 'micro-eigenschappen' van datgene waarvoor het symbool staat, symbolen zijn vectoren van subsymbolen. Uitgaande van dit idee zijn er diverse manieren waarop knopen, groepen van knopen, symbolen en subsymbolen zich kunnen verhouden. *In abstracto* kunnen drie mogelijkheden worden onderscheiden:

1. Knopen zijn symbolen.
2. Knopen zijn subsymbolen, en groepen van knopen vormen symbolen.
3. Groepen van knopen zijn subsymbolen.

De eerste en de laatste mogelijkheid staan voor respectievelijk extreem locale en extreem gedistribueerde representatie, terwijl mogelijkheid (2) een middenweg vertegenwoordigt.

Indien symbolen *locaal* gerepresenteerd zijn, zoals in mogelijkheid (1), voert een netwerk automatisch ook berekeningen uit op symbolen; de berekeningen in connectionistische modellen grijpen immers aan op individuele knopen, *ergo* op de locale symbolen. In het tweede geval, wanneer de activatietoestanden van de afzonderlijke knopen kunnen worden geïnterpreteerd als 'micro-eigenschappen', zijn het de *sub*sym-

bolen die locaal worden gerepresenteerd en dus tevens worden berekend door het netwerk. Ook in dit geval kan moeilijk worden volgehouden dat symbolen niet worden berekend. De betekenis van de symbolen is immers een functie van die van de subsymbolen; als het netwerk subsymbolen berekent, berekent het automatisch ook de daaruit samengestelde symbolen.

Het meest extreme geval van gedistribueerde representatie wordt gevormd door mogelijkheid (3), waarbij subsymbolen, samenvallend met de laagste trap van cognitief zinvolle interpretatie, worden gevormd door *groepen* van knopen. De activatietoestanden van de afzonderlijke knopen die deel uitmaken van een subsymbool kunnen nu zelf niet meer cognitief zinvol worden geïnterpreteerd als representaties van micro-eigenschappen. Ook het onderverdelen van micro-eigenschappen in nano- of pico-eigenschappen biedt geen principiële uitweg. Voor de nano- en pico-eigenschappen keert de vraag naar de relatie tussen knoop en eigenschap gewoon terug, zodat een oneindige regressie dreigt. Zelfs indien de afzonderlijke knopen zich niet cognitief zinvol laten interpreteren, kan moeilijk worden volgehouden dat de gedistribueerde subsymbolen en symbolen niet worden berekend. Uit de algoritmen voor de berekening van de activatietoestanden van de afzonderlijke knopen volgt immers onmiddellijk ook een algoritme voor de berekening van de vector van die toestanden; het enige dat verandert is de wijze van notatie. Het netwerk berekent dus niet alleen de activatiewaarden van de *afzonderlijke knopen*, maar tevens de waarden van de daaruit *samengestelde vectoren:* het netwerk berekent subsymbolen, *ergo* berekent het ook de daaruit samengestelde symbolen.

Wij kunnen op grond van het bovenstaande constateren dat het connectionisme zich ook onder de microscoop voordoet als een vorm van computationalisme, in de zin dat de door netwerken uitgevoerde berekeningen zich niet alleen *onder*, maar ook *op* het niveau van symbolen afspelen. Volgt hieruit nu dat connectionistische modellen slechts een implementatie zijn van een orthodoxe, syntactische *language of thought*, net zoals een programma in Prolog wordt geïmplementeerd in de machinetaal van een computer? Als dat het geval is, zou het connectionisme cognitief en filosofisch gezien niets nieuws te bieden hebben: op welke hardware en onder welk systeem zijn programma draait laat de gebruiker ervan immers ook Siberisch. Aan dit beeld van het connectionisme is de volgende paragraaf gewijd. Hoewel het idee van een 'implementationeel connectionisme' zeker tot de mogelijkheden behoort, merken wij hier al op dat het niet noodzakelijk juist hoeft te zijn. Een van de alternatieven zou kunnen zijn dat het connectionisme, met zijn abstracte mathematische beschrijving van een nieuwe soort hardware die sterk doet denken aan die van de menselijke hersenen, de fascinerende mogelijkheid biedt om de cognitieve programmatuur betrekkelijk direct op de hardware te projecteren. Het is

daarbij alleszins voorstelbaar dat de grenzen en mogelijkheden van netwerkmodellen nieuwe inzichten in het programmeren van cognitieve functies met zich meebrengen, inzichten die ons noodzaken het gangbare beeld van deze cognitieve functies te verfijnen en/of bij te stellen. Zoals boven al werd vermeld, is dit het tweede aspect van Smolensky's visie op het connectionisme: connectionistische modellen geven *preciezer* weer wat traditionele modellen slechts *bij benadering* beschrijven. Op deze mogelijkheid komen wij terug in de slotparagraaf, handelend over diverse varianten van eliminatief connectionisme.

Wat gesymboliseerd wordt is niet berekend: implementationisme

Zoals boven al werd vermeld, is de tweede mogelijke interpretatie van het connectionisme, evenals het zojuist behandelde subsymbolisme, gebaseerd op de premisse dat er weliswaar berekeningen plaatsvinden in netwerken, maar dat deze berekeningen zich afspelen *onder* het niveau van de symbolen. De conclusie die hieruit wordt getrokken is ditmaal echter volkomen tegenovergesteld. Symbolische eigenschappen zijn nu niet secundair ten opzichte van de elementaire berekeningen op subsymbolisch niveau, maar juist omgekeerd: het subsymbolische is secundair ten opzichte van het symbolische. Deze interpretatie van het connectionisme wordt verdedigd door onder anderen Fodor en Pylyshyn.[14] Aangezien de kern van cognitie wordt gevormd door symbolen, zo redeneren zij, zijn berekeningen die zich onder het niveau van symbolen afspelen niet relevant voor de verklaring van cognitieve verschijnselen. De rekennetwerken van hel connectionisme kunnen daarom hooguit worden beschouwd als een beschrijving van de *hardware* waarin een cognitief programma is geïmplementeerd, een beschrijving die cognitief gezien irrelevant is.

Fodor en Pylyshyn argumenteren dat de knopen en connecties niet cognitief relevant kunnen zijn omdat zij twee essentiële eigenschappen missen: ten eerste hebben zij *geen combinatorische semantiek*, en ten tweede zijn de processen die zich in een netwerk afspelen *niet structuurgevoelig*. Bij de bespreking van het orthodox computationalisme in een eerdere paragraaf is het bezit van een combinatorische structuur naar voren gekomen als een van de voornaamste kenmerken van mentale representaties. Complexe symbolen zijn volgens bepaalde regels samengesteld uit elementaire bouwstenen, net zoals zinnen in een taal zijn samengesteld uit woorden. De betekenis van een complexe representatie is een functie van die van de bouwstenen. Mentale processen, d.w.z. processen waarbij een bepaalde cognitieve functie wordt berekend, zijn bovendien gevoelig voor de structuur van de te verwerken symbolen. Het symbool '$A \rightarrow B$' zal onder omstandigheden een andere ftinctiewaarde opleveren dan het symbool '$B \rightarrow A$', omdat de beide

symbolen op verschillende wijze zijn opgebouwd uit de bouwstenen 'A', '→' en 'B'. Volgens het computationalisme, zo hebben wij in figuur 2 gezien, wordt deze combinatorische structuur op semantisch niveau (het niveau van de interpretatie van de mentale symbolen) weerspiegeld op syntactisch niveau (het niveau van de berekening van de symbolen) en op fysisch niveau (het niveau van de onderliggende causale processen).

Het hebben van een combinatorische semantiek en syntaxis brengt een aantal typische eigenschappen met zich mee die van belang zijn voor het verklaren van cognitieve verschijnselen. Tot de voornaamste daarvan behoren *systematiciteit* (als het model kan representeren dat A→B, dan kan het ook representeren dat B→A, aangezien beide complexe representaties zijn samengesteld uit dezelfde elementaire representaties 'A', '→' en 'B'), *productiviteit* (uit een beperkt aantal bouwstenen kan het model een onbeperkt aantal samengestelde representaties vormen), en semantische en inferentiële *coherentie* (als het model weet dat P en dat P→Q, dan zal het niet besluiten dat niet-P of dat niet-Q). Deze eigenschappen worden elders in deze bundel behandeld, zodat wij er hier niet verder op zullen ingaan. In plaats daarvan concentreren wij ons op de fundamentele eigenschap waaruit de overige voortvloeien: het bezit van een combinatorische semantiek en dito syntaxis.[15]

Volgens Fodor en Pylyshyn hebben connectionistische modellen, anders dan traditionele symbolistische modellen, geen combinatorische semantiek en syntaxis omdat zij alleen *causale* relaties tussen netwerkknopen aannemen. Indien bekend is hoe de inhibities en activaties in het netwerk lopen, weten wij alles wat nodig is om de werking van het netwerk te begrijpen. Klassieke theorieën, daarentegen, kennen naast causale relaties tussen de fysische symbolen ook nog tal van andere structurele relaties op syntactisch en semantisch niveau. In bovenstaande schets hebben wij daarvan een belangrijk voorbeeld gezien, nl. de relatie van samengesteldheid uit bouwstenen. Aangezien er volgens Fodor en Pylyshyn tussen de knopen in netwerken geen ruimte is voor relaties zoals samengesteldheid, kunnen connectionistische modellen nimmer een adequaat beeld geven van cognitie. De processen die zich in een netwerk afspelen zijn louter causaal van aard; zij zijn met name niet gevoelig voor de interne structuur van mentale representaties, aangezien de knopen waarop zij aangrijpen helemaal geen interne structuur hebben.

Men zou nu echter kunnen opperen dat er weliswaar tussen de afzonderlijke knopen van een netwerk alleen causale relaties bestaan, maar dat er russen de *interpretaties* van de knopen en groepen van knopen als representatie van bepaalde objecten of (micro-)eigenschappen, zoals die tot uitdrukking komen in de *labels* van de knopen (bijvoorbeeld het label 'H+' in figuur 3), ook andere relaties mogelijk zijn,

onder meer die van samengesteldheid. Zo stelt Smolensky in zijn repliek op Fodor en Pylyshyn dat de relatie van samengesteldheid in connectionistische modellen terug te vinden is als een relatie tussen vectoren en deelvectoren van activatiepatronen.[16]

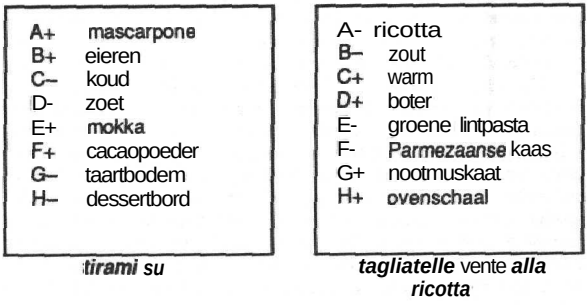
In de vorige paragraaf zagen wij dat vectoren en deelvectoren kunnen worden beschouwd als de dragers van semantische waarde. In het netwerk voor waarneming van de Necker-kubus (figuur 3), bijvoorbeeld, kunnen de activatieverdelingen {A+, C-, B+, D-, E+, G-, F+, H-} en {A-, C+, B-, D+, E-, G+, F-, H+} ieder worden weergegeven als een vector in een 16-dimensionale ruimte. De twee mogelijke ruimtelijke standen van de kubus kunnen nu onder meer worden weergeven als een samenstelling van deze vectoren, laten wij zeggen als (1, 0) en (0, 1). De interpretatie van deze vectoren is een functie van die van de deelvectoren, die op haar beurt wordt bepaald door de interpretatie van de afzonderlijke knopen, d.w.z. door *de labels* van de knopen, die als evenzovele deelvectoren in de 16-dimensionale ruimte fungeren. Nu is het weliswaar zo dat de verdeling van activa lie over het netwerk op zuiver causale wijze tot stand is gekomen, en dat tussen de afzonderlijke knopen enkel causale relaties bestaan, maar toch kent het netwerk op *semantisch* niveau een cognitief relevante, combinatorische structuur. Bovendien spelen zich in het netwerk, *semantisch* gezien, wel degelijk structuurgevoelige processen af. Wanneer bijvoorbeeld aanvankelijk alleen de knopen {A+}, {B+} en (H+) actief zijn, kan het daaropvolgende proces worden beschreven als het minimaliseren van het aantal inconsistente en het maximaliseren van het aantal consistente hypothesen over de ruimtelijke stand van de hoekpunten en van de kubus als geheel. Gedurende dit proces worden structurele relaties tussen de interpretaties (labels) van de knopen onderzocht en beoordeeld op hun consistentie; de combinatie {A+, B+, F+, H+) zal worden verworpen, de combinatie {A+, B+, F+, H-) zal worden aanvaard, enzovoorts.

Een bezwaar dat volgens Fodor en Pylyshyn aan een dergelijke redenering kleeft is dat de labels waarmee de knopen in een netwerk door de onderzoeker worden gemerkt door het netwerk *zelf* niet gelezen kunnen worden. Het netwerk zelf heeft geen boodschap aan de labels of vectoren; het kan in zijn berekeningen enkel rekening houden met de causale, inhiberende of stimulerende relaties tussen de afzonderlijke knopen.

> Strictly speaking, the labels play no role at all in determining the operation of a Connectionist machine; in particular, the operation of the machine is unaffected by the syntactic and semantic relations that hold among the expressions that are used as labels. To put this another way, the node labels in a Connectionist machine are not part of the causal structure of the machine.[17]

Al is deze objectie op zich helemaal juist, zij lijkt ons hier echter misplaatst te zijn, Fodor en Pylyshyn doen het voorkomen alsof de situatie in conventionele computationele machines anders zou zijn, d.w.z. alsof de interpretatie van de causale structuur van conventionele machines wèl een deel zou zijn van die causale structuur zelf. Daarmee gaan zij voorbij aan het door het computationalisme gemaakte onderscheid tussen enerzijds de causale relaties op fysisch en syntactisch niveau, en anderzijds de interpretatie van deze relaties in termen van cognitieve functies (zie figuur 2). Misschien is deze vergissing ingegeven door een andere stelling van het computationalisme, nl. dat relaties op semantisch niveau worden weerspiegeld op syntactisch en fysisch niveau. Hetgeen wordt weerspiegeld is echter nog geen onderdeel van de spiegel zelf! In conventionele machines zijn de interpretaties van de symbolen net zomin een onderdeel van de causale structuur als in connectionistische machines; de causale structuur is alleen zodanig dat bepaalde causale processen systematisch een bepaalde interpretatie toelaten.

Dat laatste is ook in connectionistische netwerken het geval. In het netwerk voor de waarneming van de kubus van Necker (figuur 3), bijvoorbeeld, zijn de causale verbindingen tussen de knopen zodanig ingericht dat zij de structurele, ruimtelijke relaties tussen de door de knopen gerepresenteerde hoekpunten weerspiegelen. Daarmee is echter nog niet gezegd dat de causale structuur van het netwerk de semantiek ervan *volledig vastlegt*, laat staan dat de semantiek een *onderdeel* van de causale structuur zou zijn. Een eenvoudig voorbeeld moge volstaan om dit punt toe te lichten. Precies hetzelfde netwerk van figuur 3, d.w.z. precies dezelfde causale structuur, kan ook worden gebruikt voor de herkenning van twee eenvoudige gerechten uit de Italiaanse keuken. Als het netwerk op de goede manier wordt verbonden met de zintuigen (een veronderstelling die ook in figuur 3 voor lief is genomen), kunnen de afzonderlijke knopen worden geïnterpreteerd als (bijvoorbeeld) representaties van diverse eigenschappen van respectievelijk het zoete nagerecht *tirami su* (een smakelijk taartje) en het hartige pastagerecht *tagliatelle verde alla ricotta* (gegratineerde groene lintpasta met Italiaanse kwark). Het netwerk fungeert nu als een Italiaanse *food processor*, zogezegd. In figuur 4 is een van de mogelijke verzamelingen nieuwe interpretaties van de knopen afgebeeld. Merk op dat het *netwerk op zich* onveranderd blijft. Het enige dat verandert is de manier waarop het netwerk wordt verbonden met de omgeving, hetgeen hier tot uitdrukking komt in de labels van de knopen. De knopen, hun drempelwaarden, de connecties en hun gewichten blijven ongewijzigd. Ontvangt knoop C+ een positief signaal, dan betekent dat ditmaal echter niet dat hoekpunt C van de kubus waarschijnlijk naar voren wijst, maar dat wij waarschijnlijk te maken hebben met een warm gerecht. *C+* zendt een inhiberende prikkel naar C-, wat ditmaal



Figuur 4: Nieuwe labels voor oude knopen. Een Italiaanse *food processor*.

moet worden begrepen als een verzwakking van de hypothese dat wij te maken hebben met een koud gerecht, enzovoorts. Uiteindelijk komt het netwerk terecht in een stabiele toestand, ofwel in {A+, C-, B+, D-, E+, G-, F+, H-), wat een representatie van *tirami su* is, ofwel in {A-, C+, B-, D+, E-, G+, F-, H+}, wat een representatie van *tagliatelle verde alla ricotta* is.[18]

Een connectionistisch netwerk als dat van figuur 3 heeft geen *intrinsieke* interpretatie. De betekenis van *de* configuraties van activiteit van de knopen wordt niet (uitsluitend) bepaald door de interne structuur van het netwerk, maar hangt (mede) af van de wijze waarop het netwerk met zijn omgeving is verbonden. Zoals uit figuur 4 blijkt, laat een en dezelfde structuur meerdere interpretaties toe.

Wat voor de Italiaanse *fooi processor* geldt, geldt ook voor conventionele computationele architecturen. In een kritiek op de zogenaamde procedurele semantiek geeft Fodor zelf het voorbeeld van een symbolistische computer die met een en hetzelfde gecompileerde programma de ene keer een simulatie van de Zesdaagse Oorlog doorrekent, en de andere keer een partij schaak naspeelt." *De* ene keer kunnen bepaalde symbolen in het program stelselmatig worden geïnterpreteerd als infanteriedivisies en tankbataljons, als Moshe Dayan, bommen en oorlog; de andere keer kunnen fysiek dezelfde symbolen worden geïnterpreteerd als pionnen en torens, de witte koning, rokeren en schaken. Let wel: deze verschillende interpretaties zullen ongetwijfeld terug te vinden zijn in de namen van de procedures en datastructuren waarvan de in een hogere programmeertaal (laten wij zeggen: in Lisp) geschreven programma's zich bedienen. In de *machinetaal* vinden wij deze interpretaties evenwel niet terug; de in Lisp onderscheiden programma's, eenmaal in machinetaal gecompileerd, zijn in die zin identiek. Pregnant uitgedrukt: de machine *zelf* kan de interpretaties van haar syntactische en causale processen niet lezen. De semantiek van de onderliggende processen komt niet uit de machine zelf, maar uit haar interactie met de omgeving. In het geval van een (symbolistische dan wel connectionistische) computer, wordt de semantiek in het bijzonder bepaald door het gebruik dat de ontwerper of de programmeur van de machine maken, d.w.z. door hun interpretatie van de input en output.

In de recente 'psychosemantische' literatuur wordt door steeds meer schrijvers het semantisch belang van *de* interactie tussen de computationele machine (computer dan wel cognitief organisme) en haar omgeving onderkend. Volgens deze zogenaamde causale theorieën van mentale representatie, waarmee ook Fodor zich soms vereenzelvigt, is de betekenis van mentale symbolen een functie van de causale relatie tussen het kenapparaat en objecten en eigenschappen in zijn omgeving.[20] Het spreekt voor zich dat deze causale theorieën gelijkelijk opgaan voor symbolistische en connectionistische machines; op het punt van hun *semantiek* kunnen beide soorten van machines dientenge-

volge meer en meer worden gelijkgesteld. Voor de in deze paragraaf besproken implementationistische interpretatie van het connectionisme door Fodor en Pylyshyn betekent een en ander dat netwerkmodellen niet zonder meer als cognitief irrelevant hoeven te worden beschouwd. Anders dan Fodor en Pylyshyn betogen, zijn netwerken *qua semantiek* niet zonder meer ongeschikt voor de verklaring van cognitieve verschijnselen.

De berekende symbolen vormen geen cognitieve functie: eliminisme

Voor de derde en laatste interpretatie van het connectionisme werpen wij nogmaals een blik op figuur 2 hierboven. Volgens het computationalisme wordt in het kenapparaat op syntactisch niveau een functie Φ van a naar ß berekend, die op cognitief niveau kan worden geïnterpreteerd als de berekening van een cognitieve functie V van de interpretatie van a naar de interpretatie van p. De derde opvatting van het connectionisme bestaat nu hierin, dat de processen die zich in netwerken van knopen afspelen inderdaad kunnen worden beschouwd als de berekening van een functie Φ van a naar β, en dat a en p inderdaad kunnen worden geïnterpreteerd als symbolen, maar dat er op cognitief niveau geen functie Ψ tussen de interpretaties van a en ß bestaat. In feite wordt de bovenste pijl in figuur 2 geëlimineerd: de berekende symbolen hangen niet samen als de argumenten en functiewaarden van de ons vertrouwde cognitieve functies zoals waarnemen, zich herinneren, redeneren en dergelijke meer.

Deze opvatting, die bekend staat als *eliminatief materialisme* en een fervent pleitbezorger heeft in de persoon van Paul Churchland[21] ontkent de werkelijkheidswaarde van de traditionele cognitieve functies. Zij zouden berusten op een verkeerde voorstelling van het mentale en zouden daarom voor cognitief-wetenschappelijke doeleinden irrelevant zijn. De verhouding tussen netwerken en traditionele cognitieve functies zou volgens de eliminist ongeveer te vergelijken zijn met die tussen het heliocentrisme en het geocentrisme in de astronomie. In het dagelijks leven kunnen wij zonder enig probleem zeggen dat de zon opkomt en dat zij ondergaat; dat is een doeltreffende beschrijving van een astronomisch verschijnsel. Het zou echter te ver gaan om te zeggen dat de zon echt opkomt, in de zin dat zij zich boven de horizon verheft omdat zij om de aarde draait. De geocentrische hemelmechanica waaraan het spraakgebruik is ontleend heeft heden ten dage afgedaan; de ware oorzaak van het verschijnsel van zonsopgang wordt beschreven door het heliocentrisme: zonsopgang is in feite 'aardsondergang'. Analoog hieraan is het volgens de eliminist weliswaar correct om bepaalde toestanden van netwerken te interpreteren als representaties van (bijvoorbeeld) premissen en conclusies, maar het zou te

ver gaan om te zeggen dat de door het netwerk berekende functies tussen deze symbolen ook echt redeneringen zijn. *De* 'volkspsychologie' waaraan dit spraakgebruik is ontleend is een primitieve en onjuiste theorie; de *ware* functies die worden berekend worden beschreven door de onderliggende connectionistische differentiaalvergelijkingen en door de neurofysiologie.[22]

De eliminatieve interpretatie van het connectionisme kan op diverse punten worden aangevallen. Aangezien elders in deze bundel de argumenten voor en tegen het eliminisme al uitvoerig ter sprake worden gebracht, zullen wij ons hier beperken tot de hoofdlijnen. Het voornaamste probleem waarmee deze opvatting te kampen heeft, is dat tegelijk met de relatie tussen symbool en cognitieve functie, ook de relatie russen computationalisme en cognitie op losse schroeven wordt gezet. Wanneer het computationalisme niet dient ter verklaring van de bekende cognitieve functies, wat is dan zijn functie? In feite berooft deze derde interpretatie het computationalisme en de cognitiewetenschap in één klap van al hun explananda; het is aan de eliminist om als het ware uit het niets een volledige verzameling nieuwe explananda te voorschijn te toveren. Daarbij is het echter nog maar de vraag in welke zin een cognitiewetenschap die niet de ons vertrouwde cognitieve verschijnselen verklaart, überhaupt nog een wetenschap van *cognitie* kan worden genoemd. Het fenomeen 'cognitie', zo zou men immers kunnen redeneren, wordt gedefinieerd en afgebakend in termen van de traditionele, 'volkspsychologische' categorieën en begrippen zoals waarnemen, zich herinneren, redeneren, enzovoorts; wanneer men, zoals de eliminist voorstelt, ontkent dat deze categorieën corresponderen met iets in de werkelijkheid, ontkent men in feite de realiteit van cognitie zelf.[23]

Op grond van deze overwegingen lijkt een *radicule* eliminatie van traditionele cognitieve functies nauwelijks een reële mogelijkheid te zijn. Dit betekent echter niet dat eliminatie helemaal uit den boze is. Er is een gematigd alternatief denkbaar, bestaande in een gedeeltelijke, stap-voor-stap verlopende eliminatie en revisie van de traditionele cognitieve functies. Zonder in één klap alle cognitieve verschijnselen overboord te zetten, zou de cognitiewetenschap de cognitieve functies tentatief en heuristisch kunnen overnemen teneinde te onderzoeken welke syntactische en causale processen voor de berekening van deze functies verantwoordelijk zouden kunnen zijn. Wanneer blijkt dat het traditionele beeld van een bepaalde functie te grof of zelfs onjuist is, moet de functie worden bijgesteld, verfijnd, of zelfs worden geëlimineerd. Deze geleidelijke revisie en eliminatie vermijdt het boven gesignaleerde explanandum-probleem. De volkspsychologie en de cognitieve psychologie dragen de explananda aan, die door het connectionisme tentatief ter verklaring worden overgenomen; vervolgens worden deze explananda op grond van de connectionistische bevindingen zonodig bijgesteld of verduidelijkt.

Een soortgelijke werkverdeling tusen connectionisme en orthodoxe cognitieve psychologie lijkt ook door Smolensky te worden voorgestaan. In een vergelijking met de relatie tussen de klassieke mechanica en de quantummechanica wijst hij erop dat de klassieke mechanica niet zonder meer wordt geïmplementeerd noch wordt geëlimineerd door de quantummechanica, maar deze veeleer *benadert:* de quantumthéorie geeft preciezere verklaringen voor wat de klassieke mechanica aan explananda schetst. De laatste blijft daarbij van essentieel theoretisch belang, niet alleen ter (approximatieve) verklaring van de klassieke verschijnselen, waarvoor de quantummechanica ongeschikt is, maar ook "to provide the guidance necessary to discover the quantum principles in the first place".[24]

Toegepast op de verhouding tussen klassieke cognitieve functies en connectionistische modellen, zou men op grond van deze vergelijking kunnen verdedigen dat netwerkmodellen in feite ergens moeten worden ingeschaald *tussen* het niveau van de hardware en dat van de software, tussen de neurofysiologie en de traditionele cognitieve psychologie. In figuur 5 is schematisch weergegeven hoe in zo'n geval de diverse niveaus van analyse van een computationeel systeem zich verhouden. De cognitieve psychologie beschrijft *welke* cognitieve functies er worden berekend; het connectionistisch model beschrijft *hoe* deze berekeningen precies worden uitgevoerd; en de neurobiologie beschrijft hoe het rekenapparaat zelf gebouwd is. Het onderzoek op elk van deze niveaus is via relaties van revisie en heuristiek 'teruggekoppeld' aan zijn buren. Het intermediair verklarings- en beschrijvingsniveau van het connectionisme zou daarbij de conceptuele afstand tussen biologische roetoare en cognitieve *software* aanmerkelijk kunnen verkleinen. Het is dan ook niet ondenkbaar dat het connectionisme goede diensten kan bewijzen als een soort van 'conceptuele brug' waarlangs toenadering en samenwerking tussen neurowetenschap en cognitieve psychologie mogelijk is.

In zekere zin kan deze werkverdeling tussen connectionistische en klassieke modellen van cognitie worden beschouwd als een moderne opvolger van het traditionele functionalisme in de *philosophy of mind.* Er bestaat echter een belangrijk verschil tussen erflater en erfgenaam. Het functionalisme heeft zich steeds op het standpunt gesteld dat lagere niveaus van analyse essentieel irrelevant zijn voor het begrip van verschijnselen op hogere niveaus van analyse. Het standaardargument achter deze opvatting is dat van 'meervoudige realiseerbaarheid': elke cognitieve functie kan op talloze manieren worden gerealiseerd in de meest uiteenlopende fysische substraten; de details van de implementatie zijn daarom irrelevant voor het verklaren van de functie. De revisionistische opvatting van het connectionisme, daarentegen, laat ruimte voor een voortdurende wisselwerking tussen de diverse niveaus van analyse. Volgens sommigen betekent het connectionisme dan ook

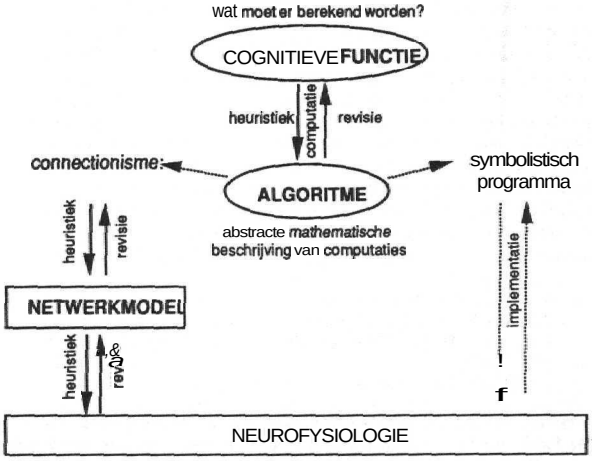
een weerlegging van het argument van meervoudige realiseerbaarheid.[25]

Het hier in het kort geschetste, speculatieve beeld van een 'revisionistisch connectionisme' zou kunnen worden gestaafd door de verdere ontwikkeling van connectionistische modellen. In dit verband is het van belang te wijzen op het werk van David Marr. Het door Marr ontwikkeld onderzoeksmodel, dat veel overeenkomst vertoont met dat in figuur 5, kan bogen op aanzienlijk empirisch succes, juist omdat het ruimte schept voor een intensieve uitwisseling van begrippen en re stricties tussen de diverse niveaus van analyse van computationele systemen.[26]

Besluit

In dit artikel is een aantal filosofische aspecten van het connectionisme de revue gepasseerd, waarbij de nadruk heeft gelegen op de semantische aspecten van netwerkmodellen van cognitie. Op grond van een toetsing van de harde kern van het connectionisme aan de kerngedachten van het orthodoxe computationalisme moet worden geconcludeerd dat, anders dan door velen in connectionistische zowel als orthodox symbolistische kringen wordt aangenomen, het niet zonder meer duidelijk is op welke essentiële punten het connectionisme van de symbolistische traditie verschilt.

Uitgaande van het idee dat er in computationele systemen een niveau van 'natuurlijke symbolen' bestaat, werd nagegaan hoe de berekeningen die in connectionistische netwerken worden uitgevoerd zich tot deze natuurlijke symbolen verhouden. Daarbij is een drietal mogelijkheden besproken: 1. Wat berekend wordt zijn geen symbolen; 2. Wat gesymboliseerd wordt is niet berekend; 3. De berekende symbolen vormen geen klassieke cognitieve functies. Deze drie mogelijke interpretaties van het connectionisme, die wij respectievelijk subsymbolisch, implementationeel en eliminatief connectionisme hebben gedoopt, blijken ieder zo hun eigen problemen te hebben. Het subsymbolisme, zoals verdedigd door onder anderen Smolensky, bedient zich van een betrekkelijk onduidelijk idee van locale en gedistribueerde representaties. Pogingen om dit idee nader te preciseren blijken hoe dan ook uit te lopen op een ontkenning van de stelling dat niet *de* symbolen maar enkel de subsymbolen worden berekend in connectionistische netwerken. Het implementationisme, zoals verdedigd door Fodor en Pylyshyn, betoogt dat netwerkmodellen niet cognitief relevant kunnen zijn omdat zij de daartoe benodigde combinatorische semantiek missen. Deze opvatting blijkt bij nader onderzoek gebaseerd te zijn op het idee dat computationele systemen moeten beschikken over een intrinsieke semantiek, een vooronderstelling die op zijn minst



Figuur 5: Diverse niveaus van analyse van computationele systemen.

controversieel moet worden genoemd. Het eliminisme tenslotte, zoals verdedigd door Churchland, ontkent de realiteit van (klassieke) cognitieve functies; de door connectionistische netwerken berekende symbolen fungeren niet als argumenten en functiewaarden van ons bekende cognitieve functies. Door het verbreken van de band tussen computatie en cognitie, berooft deze opvatting het computationalisme van zijn explanandum en de cognitiewetenschap van haar object. Het is nog maar de vraag in hoeverre een wetenschap die niet de klassieke cognitieve functies verklaart een wetenschap van cognitie kan worden genoemd.

Op diverse punten in onze discussie heeft zich, naast de besproken drie interpretaties van het connectionisme, een gematigd alternatief aangediend dat wij 'revisionistisch connectionisme' hebben gedoopt. Dit alternatief voorziet in een wisselwerking tussen traditionele cognitief-psychologische beschrijvingen van cognitieve functies, netwerkmodellen en neurofysiologische beschrijvingen van het zenuwstelsel. De relatief abstracte wiskundige beschrijving van de dynamiek van betrekkelijk 'neuronale' netwerken vervult daarbij een brugfunctie tussen cognitieve psychologie en neurofysiologie. Misschien hebben Fodor en Pylyshyn in zekere zin gelijk wanneer zij stellen dat connectionistische modellen niet cognitief relevant kunnen zijn omdat zij de doelstellingen van de klassieke cognitieve psychologie te buiten gaan.

> In fact, [connectionist models] might be viewed as *advancing the goals of* Classical information processing psychology by attempting to explain how the brain (or perhaps some idealized brain-like network) might realize the types of processes that conventional cognitive science has hypothesized.[27]

Gegeven het feit dat kennis een verschijnsel is dat gedefinieerd is aan de hand van het voorbeeld van de met hersenen begaafde mens, zou men dit citaat echter ook zo kunnen lezen dat het wellicht tijd is de doelstellingen van de klassieke psychologie te verleggen. Misschien is de tijd gekomen dat de psychologie zich richt op de vraag *hoe de mens zijn hersenen gebruikt*. Het connectionisme zou daartoe een unieke gelegenheid kunnen bieden.

Noten

1 Voor een meer gedetailleerde inleiding tot de kernbegrippen en voornaamste theoretische aspecten van het connectionisme, zie de bijdrage elders in deze bundel van Phaf & Murre.

2 Voor een bespreking van het modulariteitsbegrip, zie de bijdragen van Meijering en Michon & Jorna. Voor een bespreking van de plasticiteit van connectionistische modellen, zie de bijdragen van Levelt en Phaf & Murre.

3 Het idee van conceptuele 'vectorruimten' is geïnspireerd op de in connectionistische kringen gangbare voorstelling van de activatietoestanden van netwerken als vectoren in een *phase space*. Zie bijv. Churchland (1986), p. 412 v.v., Churchland (1988), p. 146 v.v. en 156 v.v.

4 Voor een schets van Rumelhart & McClelland's model voor letter- en woordherkenning, zie de bijdrage van Phaf & Murre elders in deze bundel.

5 Fodor & Pylyshyn (1988), p. 30.

6 Fodor (1975), **Pylyshyn** (1984). Voor **een systematische bespreking en verdediging van het idee van een** *language of thought*, **zie ook het appendix in Fodor** (1987), p. 135 **v.v.**

7 Voor een gedetailleerde bespreking van deze verschillen tussen connectionistische en symbolistische modellen, zie de bijdrage van Levelt elders in deze bundel, alsmede Clark (1987), Smolensky (1988), Pinker & Prince (1988) en Fodor & Pylyshyn (1988).

8 Vrij naar Rumelhart & McClelland 1986, vol. II, p. 8 v.v. Het hier geschetste netwerk dient uitsluitend ter illustratie van bepaalde abstracte eigenschappen van connectionistische netwerken in het algemeen. Het maakt geen enkele aanspraak op een speciale psychologische, neurofysiologische of AI-plausibiliteit. Er zijn ook andere netwerken ontwikkeld voor de waarneming van de kubus van Necker, onder anderen door Feldman, die wèl aanspraak maken op een dergelijke plausibiliteit. De hier geïllustreerde eigenschappen van connectionistische netwerken gelden uiteraard evenzeer voor die meer realistische modellen.

9 Volledigheidshalve moet worden aangetekend dat het hier gegeven voorbeeld in werkelijkheid iets te eenvoudig is. De kans bestaat dat aldus toegeruste netwerken er niet in slagen de beste oplossing voor een cognitief probleem te vinden, maar (als een mot in een kaars) vastlopen in een oplossing die alleen locaal gezien de beste is. Veel aandacht in het connectionistisch modelleren gaat op dit moment dan ook uit naar de ontwikkeling van algoritmen om een netwerk in staat te stellen de globaal gezien beste oplossing te vinden. De basisprincipes van *constraint networks* blijven echter onverlet.

10 Zie onder meer Clark (1987), p. 13-14.

11 Smolensky (1987), p. 154-155.

12 Smolensky (1987), p. 152. Zie ook Smolensy (1988), p. 3, 6, 7-8 en 16-17, waar deze 'subsymbolic hypothesis' nader wordt uitgewerkt. Een soortgelijke positie wordt ingenomen door onder anderen Hofstadter; zie diens *Waking up from the Boolean dream*, in Hofstadter (1985), p. 631 v.v.

13 Smolensky (1988), p. 7.

14 Fodor & Pylyshyn (1988). Zie ook Pylyshyn (1984), hoofdstukken 3 en 7, en Fodor (1987), p. 135 v.v. (appendix).

15 Wij verwijzen de lezer naar de bijdrage van Levelt elders in deze bundel. Zie verder Fodor & Pylyshyn (1988), p. 12 v.v., p. 33 v.v. Vgl. ook Fodor (1987), p. 135 v.v.

16 Smolensky (1987, 1988).

17 Fodor & Pylyshyn (1988), p. 17.

18 Volledigheidshalve vermelden wij dat de hier afgebeelde lijsten van kenmerken niet bedoeld zijn als volledige recepten. De lezer zal, geïnteresseerd geraakt in de Italiaanse keuken, gemakkelijk zelf de ontbrekende ingrediënten kunnen aanvullen.

19 Fodor (1981), p. 207 v. v.
20 Zie bijv. Fodor (1987). Voor een kritische bespreking van enkele der voornaamste causale theorieën, *zie* Sleutels (1989),
21 Zie onder meer Churchland (1981), en (1988), p. 43 v.v., alsmede het encyclopedische werk *Neurophilosophy* van Paul Churchland's echtgenote Patricia (1986).
22 Zie onder meer Churchland (1981), p 84 v.v.
23 Voor een uitgebreide kritiek op het eliminatief materialisme, zie Sleutels (1988).
24 Smolensky (1987), p. 154-155.
25 Zo onder anderen Thagard (1986).
26 Zie Marr (1982) of, voor een beknopte samenvatting van Marr's onderzoeksmodel, de bijdrage van Bürge in Garfield (1987). Op het model van Marr is een succesvol onderzoeksprogramma van *natural computation* geënt; zie Richards (1988). Enkele belangrijke aspecten van Marr's model worden besproken in Sleutels (1988) en (1989).
27 Fodor & Pylyshyn (1988), p. 65; cursivering JS & BG.

Literatuur

Churchland, P.M., Eliminative materialism and the propositional attitudes. *Journal of Philosophy 78*, 1981, 67-91.

Churchland, P.M., *Matter and consciousness: A contemporary introduction to the philosophy of mind.* Cambridge, Mass.: MIT Press 1988.

Churchland, P.S., *Neurophilosophy: Toward a unified science of the mind/brain.* Cambridge, Mass.: MIT Press 1986.

Clark, A., Connectionism and cognitive science. In: J. Hallam & C. Mellish (red), *Advances in artificial intelligence.* (Proceedings of the 1987 AISB Conference, University of Edinburgh, 6-10 April 1987.) Chichester: John Wiley & Sons 1987.

Fodor, J.A., *The language of thought.* New York: Crowell 1975.

Fodor, J.A., *Representations: Philosophical essays on the foundations of cognitive science.* Cambridge, Mass.: MIT Press 1981.

Fodor, J.A., *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, Mass.: MIT Press 1987.

Fodor, J.A. & Pylyshyn, Z.W., Connectionism and cognitive architecture. *Cognition 2g*, 1988, 3-72.

Gardner, H., *The mind's new science: A history of the cognitive révolution.* New York: Basic Books 1985.

Garfield, J.L. (red), *Modularity in knowledge representation and natural-language processing.* Cambridge, Mass.: MIT Press 1987.

Hofstadter, D.R., *Metamagical themas: Questing for the essence of mind and pattern.* New York: Basic Books 1985.

Marr, D., *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman & Co 1982.

Pinker, S. & Prince, A., On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition 28*, 1988, 73-193.

Pylyshyn, Z.W., *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, Mass.: MIT Press 1984.

Richards, W. (red), *Natural computation*. Cambridge, Mass.: MIT Press 1988.

Rumelhart, D.E. & McClelland, J.L. (red), *Parallel distributed processing: Explorations in the microstructure ofcognition.* Cambridge, Mass.: MIT Press 1986,

Sleutels, J.J.M., Eliminatief materialisme en de autonomie van de bottom-up benadering. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte80*, 1988, 25-45.

Sleutels, J.J.M., Natuurlijke teleologie en het probleem van misrepresentatie in de fysicalistische philosophy of mind. *Nijmegen Studies in the Philosophy of Nature and lts Sciences W*, 1989 (in druk).

Smolensky, P., The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *The Southern* journal *of Philosophy 26*, 1987, Supplement, 137-161.

Smolensky, **P., On the proper** treatment of connectionism. *The Behavioral and Brain Sciences l 1*, 1988, 1-74.

Thagard, P., Parallel computation and the mind-body problem. *Cognitive Science 10*, 1986,301-318.