# De mens als computer

Gert-Jan Lokhorst

27 april 2000 (VERBETERDE VERSIE)

## 1 De mens als machine

De mens is in de afgelopen drie eeuwen vaak vergeleken met allerlei soorten machines.

In de achttiende eeuw was de *klokmetafoor* tamelijk populair; psychologische termen als 'drijfveer', 'van slag raken' en 'opgewonden zijn' herinneren hier nog aan [Vroon and Draaisma, 1985].

In de negentiende eeuw overheerste de *stoommachine-metafoor*. De psychologie van Freud wordt wel als een uitgewerkte versie van deze metafoor beschouwd [Russelmann, 1983]. Ook uitdrukkingen als 'uitlaatkleppen', 'stoom afblazen' en 'iemand opstoken' zijn eraan te danken. De stoommachine-metafoor wordt nog steeds serieus genomen. Zo pleegt men de menselijke geest in de nieuwe 'dynamische' school in de cognitieve wetenschappen bij voorkeur te vergelijken met James Watts centrifugale regulateur (1788), het apparaat dat ervoor zorgt dat een stoommachine op een constant snelheid werkt [van Gelder, 1995, 1998].

De laatste vijftig jaar komt men de metafoor van de *seriële digitale computer* vaak tegen. Een PC is een voorbeeld van zo'n computer. Hij is *serieel* omdat de centrale processor slechts één berekening tegelijk kan uitvoeren; hij is *digitaal* omdat hij alleen met gehele getallen kan omgaan. Het voorbeeld bij uitstek van een seriële digitale computer is de Turing-machine, waarover aanstonds meer. De seriële digitale computer metafoor is op verschillende manieren op mensen toegepast. Zo beschouwen sommigen de gehele mens als een computer van dit type, terwijl anderen menen dat de afzonderlijke zenuwcellen op deze manier beschreven kunnen worden.

In het onderstaande zal ik de vraag bespreken in hoeverre men de mens, of onderdelen van de mens, inderdaad als seriële digitale computer mag opvatten. Ik zal laten zien dat deze vergelijking maar een beperkte waarde heeft. Het valt niet uit te sluiten dat de mens in computationeel opzicht een veel sterker of zwakker soort machine is dan de Turing-machine.

## 2 De Turing-machine

Zoals gezegd is de Turing-machine het voorbeeld *par excellence* van een seriële digitale computer. Wat is een Turing-machine?

Een Turing-machine is een wiskundig model van een mens die rekent. Turing verzon zijn machine omdat hij Hilberts *Entscheidungsproblem* wilde oplossen [Turing, 1936]. Dit is het probleem of de wiskunde beslisbaar is. Kunnen we

het antwoord op iedere wiskundige vraag op een 'mechanische' manier berekenen? Om deze vraag te beantwoorden, moeten we natuurlijk eerst de notie van 'mechanische berekenbaarheid', ook wel 'effectieve berekenbaarheid' genoemd, preciseren. De Turing-machine levert de gewenste precisering.

Een Turing-machine bestaat uit een aan twee zijden onbegrensde band die in vakjes is onderverdeeld. Op de vakjes bevinden zich symbolen uit een eindig alfabet. Boven de band bevindt zich een lees/schrijf kop die de symbolen kan lezen en veranderen. De lees/schrijf kop wordt gecontroleerd door een eindige automaat, een apparaat dat bij iedere stap in één van een eindig aantal mogelijke toestanden verkeert. Berekeningen verlopen in stappen. Bij iedere stap wordt er een symbool op de band gelezen. Afhankelijk van dat symbool en van de toestand waarin de controle-eenheid zich bevindt, wordt er vervolgens een nieuw (mogelijk hetzelfde) symbool op het onder de lees/schrijf kop liggende vakje gezet, gaat de kop een vakje naar links of naar rechts, en gaat de controle-eenheid in een nieuwe toestand over. (Wat er precies gebeurt is afhankelijk van de manier waarop de controle-eenheid werkt; haar gedrag kan in een tabel beschreven worden.) Een berekening is af op het moment dat de machine, na een eindig aantal stappen, stopt.

Het zal duidelijk zijn dat dit een goede weergave is van de manier waarop een fantasieloze boekhouder rekent, één die zich strikt aan de regels houdt en een onbeperkte voorraad kladpapier tot zijn beschikking heeft. De these dat dit inderdaad een goed model is—met andere woorden, de these dat effectieve berekenbaarheid geïdentificeerd mag worden met Turing-machine berekenbaarheid—heet de Church-Turing these. Deze these wordt door iedereen aanvaard [Copeland, 1997b].

Het bijzondere van Turings formele model was dat hij ermee kon laten zien dat allerlei wiskundige functies op de door hem beschreven manier *niet* berekend kunnen worden. Een voorbeeld van een onbeslisbaar probleem is het zogenaamde 'stop-probleem': zal Turing-machine nummer X met input Y op de band ooit stoppen? Deze vraag kan over het algemeen niet mechanisch beantwoord worden. Hiermee was Hilberts *Entscheidungsproblem* opgelost. Het antwoord op de vraag die Hilbert stelde luidt 'nee'.

De Turing-machine is een wiskundige abstractie. De theorie die ervoor opgaat geldt echter ook voor alle later vervaardigde electronische seriële digitale computers. Ook zij hebben een geheugen waaruit ze lezen en waarin ze schrijven, werken met symbolen uit een eindig alfabet (rijtjes eentjes en nulletjes) en voeren slechts één instructie tegelijk uit, op een in de centrale processor ingebakken manier. Computers rekenen op dezelfde manier als waarop fantasieloze boekhouders dat doen en hebben dezelfde, door Turing voor het eerst onderkende, beperkingen. Dit geldt ook overigens voor alle DNA-, RNA- en quantum-computers waaraan tegenwoordig gewerkt wordt.

Hoewel de Turing-machine niet alles kan, is het toch een bijzonder krachtig apparaat. Turing meende zelf dat haar vermogens niet voor die van de mens zouden hoeven onder te doen. Hij voorspelde dat er tegen het jaar 2000 Turing-machines zouden zijn gebouwd waarvan het linguïstische gedrag niet van dat van een mens te onderscheiden zouden zijn [Turing, 1950]. Deze voorspelling is niet uitgekomen. De zogeheten Turing test (is machine X louter op grond van linguïstische input/output als machine te herkennen?) wordt sinds 1991 ieder jaar in de Loebner-competitie met allerlei soorten op verschillende manieren geprogrammeerde computers uitgevoerd. Geen enkel team heeft de hoofdprijs van

honderdduizend dollar tot dusver in de wacht gesleept. Simpele programma's die niets anders doen dan hun interviewers uitschelden maken soms nog de menselijkste indruk [Moor, 1998b].

## 3 De mens als Turing-machine

Het idee van de Turing-machine heeft men op verschillende manieren gebruikt in de psychologie en de filosofie. Sommige toepassingen zijn tamelijk onschuldig. Zo probeert men wel om computermodellen te maken van verschillende aspekten van de informatieverwerking in de mens. Hier lijkt weinig op tegen te zijn. Als men computermodellen mag maken van de luchtstromingen in de atmosfeer, de groei van planten en de nationale economie, waarom zou men dan geen computermodellen mogen maken van wat er bijvoorbeeld in het visuele systeem gebeurt? De praktijk zal dan natuurlijk nog wel moeten uitwijzen of de modellen adequaat zijn.

Anderen gaan verder. Zo heeft de filosoof Dennett een pretentieus boek geschreven waarin hij beweert dat ons bewustzijn met een Turing-machine te vergelijken is [Dennett, 1991]. Dit is natuurlijk onzinnig, al is het maar omdat één minuut introspectie ons al leert dat de inhoud van ons bewustzijn niet in duidelijk van elkaar stapjes aan ons voorbijtrekt. De bewustzijnsstroom is continu, niet discreet [James, 1890, hfdst. IX].

De bekendste toepassing van de Turing-machine in de filosofie is het Turing-machine functionalisme van Putnam [Putnam, 1960, 1963, 1964, 1967a,b, 1969, 1973]. Volgens deze opvatting zijn mentale toestanden zoals het hebben van jeuk in de linkerpink of het denken aan de ondergang van de *Titanic* te vergelijken met de logische of functionele toestanden waarin Turing-machines verkeren. Putnam heeft deze opvatting om verscheidene redenen inmiddels verlaten [Putnam, 1988, 1992]. Hij vertelde me onlangs dat er bij zijn weten ook niemand anders is die haar nog verdedigt [Putnam, 1995]. Hieraan behoeven we dus verder geen aandacht te besteden.

Het Turing-machine functionalisme is in de filosofie opgevolgd door het functionalisme *tout court*. Hierin hangt men nog wel het algemene idee aan dat psychologische verschijnselen als denken, waarnemen en redeneren als ingewikkelde rekenkundige bewerkingen moeten worden gezien, maar over de details laat men zich niet meer uit. Dit algemene idee wordt ook wel 'computationalisme' genoemd [Moor, 1998a]. Een voorbeeld uit deze school is het recente boek van Rey [Rey, 1997]. De Turing-machine komt er nog wel in voor, maar speelt er geen essentiële rol meer in. Ze is niet meer dan een voorbeeld van een symboolmanipulerende machine, kennelijk het enige voorbeeld dat de auteur kent. Men kan niet zeggen dat een boek als dat van Rey nog in het teken staat van de Turing-machine metafoor.

Kan de mens nu wel of niet als Turing-machine worden opgevat? De literatuur in de *philosophy of mind* laat ons kennelijk in het ongewisse. Er is nog een ander probleem met deze literatuur. Zoals Jack Copeland onlangs heeft laten zien, blijken zelfs bekende auteurs zoals de al eerder genoemde Dennett vaak maar nauwelijks te weten waar ze het over hebben als ze termen zoals 'Turing-machine' en 'Church-Turing these' in de mond nemen [Copeland, 1997b].

Laten we daarom deze literatuur terzijde leggen en zelf trachten te achterhalen in hoeverre de mens als Turing-machine mag worden beschouwd.

Eén ding is alvast duidelijk: de Turing-machine metafoor is zeker op zijn plaats als we het hebben over de uitwendig waarneembare verrichtingen van onvermoeibare, foutloos rekenende menselijke boekhouders. De Turing-machine was immers in eerste instantie bedoeld als model van dergelijke verrichtingen. Het enige dat men er misschien tegenin zou kunnen brengen is dat echte boekhouders geen onbeperkte voorraden kladpapier hebben.

Interessanter dan dit speciale geval is echter de vraag of mensen ook *in het algemeen* als Turing-machines kunnen worden opgevat. Daar valt *a priori* waarschijnlijk weinig over te zeggen. Weliswaar hebben Lucas en Penrose op een *a priori* manier trachten aan te tonen dat mensen krachtiger zijn dan welke Turing-machine ook, maar hun argumenten zijn niet overtuigend. Ze komen uiteindelijk neer op de bewering dat mensen krachtiger zijn dan welke Turing-machine ook, maar dat is juist wat bewezen moest worden [Lucas, 1970, Penrose, 1989, 1994].

Laten we daarom de *a priori* wetenschappen verlaten en kijken naar wat de empirische wetenschappen ons op dit punt te vertellen hebben. We beginnen met theorieën over de werking van het zenuwstelsel. We zullen zien dat sommige moderne theorieën suggereren dat de hersenen krachtiger zijn dan de krachtigste seriële digitale computer die men zich kan voorstellen. Volgens deze theorieën zijn wij dus geen Turing-machines, maar super-Turing-machines!

# 4 Het zenuwstelsel

Het oudste wiskundige model van de werking van het zenuwstelsel is te danken aan McCulloch en Pitts [McCulloch and Pitts, 1943]. In hun model komt een groot aantal eenvoudige zenuwcelletjes voor, die via allerlei verbindingen met elkaar in contact staan. Cellen kunnen gestimuleerd en geïnhibeerd worden. Iedere cel heeft een bepaalde drempelwaarde. Als de som van de stimulatie verminderd met de som van de inhibitie de drempelwaarde overschrijdt is de cel actief, anders is ze inactief. Cellen hebben dus slechts twee mogelijke activatiewaarden: aan of uit. We hebben hier te maken met een binair netwerk.

In 1956 toonde Kleene aan dat McCulloch-Pitts netwerken equivalent zijn met eindige automaten, de automaten die we al tegenkwamen in onze bespreking van de controle-eenheid van de Turing-machine [Kleene, 1956]. Hiermee was dus een duidelijk beeld verkregen van het soort computers dat mensen zijn: eindige automaten.

In de filosofie duikt de these dat mensen eindige automaten zijn voor het eerst op in 1957 [George, 1957]. Burks en Nelson hebben haar later met verve verdedigd [Burks, 1973, Lugg, 1990, Nelson, 1989]. Ook Putnam heeft de mens ooit als (probabilistische) eindige automaat beschouwd [Putnam, 1967b]. Nelsons boek is de meest uitgewerkte versie van deze opvatting [Nelson, 1989]. Helaas gaat het boek op een cruciaal punt de mist in: het sleutelbegrip 'verwachting' wordt niet in termen van eindige automaten, maar in termen van Turing-machines gedefinieerd.

McCulloch-Pitts netwerken zijn niet realistisch. Het zenuwstelsel is niet binair. Gelukkig heeft men tegenwoordig veel betere wiskundige modellen.

Het bekendst zijn de zogeheten analoge recurrente netwerken. Analoog betekent in dit geval dat de activiteit van de cellen geen kwestie van alles of niets is. De eigenschappen van het netwerk (de sterkten van de verbindingen, de acti-

vatiefuncties van de cellen en dergelijke) moeten met *reële* in plaats van gehele getallen beschreven worden. Recurrent betekent dat de signalen niet alleen van de input- naar de output-laag gaan, maar ook in omgekeerde richting mogen lopen.

Wat is de rekencapaciteit van dergelijke neurale netwerken? Hoe verhouden ze zich tot de Turing-machine? Het antwoord is verrassend: ze zijn veel krachtiger! Ze zijn net zo krachtig als Turing-machines die ja/nee vragen kunnen stellen aan externe 'orakels' die over meer kennis beschikken dan zij zelf ooit zouden kunnen doen. (Ook deze Turing-machines met orakels zijn overigens al door Turing zelf bedacht [Turing, 1939].) Men spreekt dan ook wel van 'super-Turing-machines' [Siegelmann, 1993, 1995, 1998, Siegelmann and Sontag, 1994, 1995].

In het begin van de opkomst van de theorie van analoge recurrente netwerken was men weleens bevreesd dat deze netwerken te zwak zouden zijn [Levelt, 1990]. Deze angst was dus volledig ongegrond.

De meest realistische moderne modellen van het zenuwstelsel werken net zoals het echte zenuwstelsel met *actiepotentialen* [Maass and Bishop, 1999]. De cellen beïnvloeden elkaar door met verschillende frequenties naar elkaar te vuren. Deze frequenties kunnen alle mogelijke waarden binnen een bepaald interval aannemen; er zijn dus evenveel mogelijke frequenties als er reële getallen zijn tussen 0 en 1, namelijk overaftelbaar veel. Ook de sterkten van de verbindingen tussen de cellen en de manier waarop iedere cel op input reageert worden met reële getallen beschreven. Dergelijke netwerken blijken net zo krachtig te zijn als 'gewone' analoge recurrente netwerken: het zijn super-Turing-machines [Maass, 1997].

Betekent dit nu dat de mens een super-Turing-machine is, een wezen dat in principe meer kan dan welke Turing-machine ook? Dat zou niet gek zijn! Helaas is er één factor die roet in het eten gooit: ruis.

Zenuwcellen zijn niet betrouwbaar. Ze kunnen op ieder moment uitvallen; ook wordt hun werking verstoord door thermische ruis, de Brownse beweging van de moleculen tussen de zenuwcellen, radioactiviteit vanuit de omgeving, kosmische straling uit de ruimte, en wat dies meer zij. Recente resultaten suggereren dat realistische neurale netwerken van de zojuist beschreven soort waarin ruis optreedt *op hun hoogst* even krachtig zijn als eindige automaten [Maass and Orponen, 1997, Maass and Sontag, 1999]. Ze zijn dus niet krachtiger dan de McCulloch-Pitts netwerken waarmee we hierboven begonnen en zeker niet krachtiger dan de Turing-machine.

Als we de balans opmaken, zien we dat er verscheidene soorten modellen van het zenuwstelsel bestaan. Volgens sommige van deze modellen zijn de hersenen veel zwakker dan de Turing-machine, volgens andere veel krachtiger. Er zijn trouwens ook wel modellen van eindige neurale netwerken (namelijk modellen waarin de sterkten van de verbindingen alleen maar *rationele* waarden mogen aannemen) die precies even krachtig zijn als de Turing-machine [Siegelmann, 1993]. De waarheid zal wel ergens in het midden liggen, maar we weten nog niet waar.

Laten we de zaak op een fundamenteler niveau bekijken en onderzoeken wat de fysica ons te vertellen heeft.

# 5 Fysica

Sinds Newton is bijna alles in de fysica analoog, niet discreet. In een vergelijking als $F{=}ma$ staan $F$, $m$ en $a$ voor reële getallen, niet *per se* voor gehele getallen. De gehele differentiaal- en integraalrekening waar de fysica op berust speelt zich af in het domein der reële getallen, en ga zo maar door. De quantummechanica heeft deze beschouwingswijze niet achterhaald; weliswaar kunnen allerlei grootheden slechts discrete waarden aannemen, maar er blijft nog genoeg analoogs over.

Dit opent perspectieven. Turing-machines zijn in alle opzichten discreet (beschrijfbaar in termen van gehele getallen); als de natuur in laatste instantie analoog is, zijn er veel meer dynamische systemen denkbaar dan alleen maar Turing-machines. Logici, wiskundigen en fysici worden dan ook al jarenlang gefascineerd door de vraag of de thans bekende natuurwetten het bestaan van dynamische systemen met super-Turing capaciteiten toestaan, en zo ja, of zulke systemen misschien zelfs al in de natuur bestaan zonder dat wij daarvan op de hoogte zijn [Copeland, 1997a, Copeland and Sylvan, 1999, Costa and Doria, 1991, Gandy, 1980, Kreisel, 1974, 1982, Moore, 1990, 1991, 1998, Penrose, 1989, 1994, Pour-El, 1974, Pour-El and Richards, 1979, 1981, 1982, Rubel, 1985, 1989, Scarpellini, 1963, Siegelmann, 1995, Stannett, 1990, Stewart, 1991, Svozil, 1997, Vergis et al., 1986, Wolfram, 1985]. (Een recent boek over de theorie van analoge automaten in het algemeen is [Blum et al., 1997].)

Het zou een opwindende zaak zijn als er super-Turing-machines zouden bestaan, al is het maar omdat dit zou impliceren dat het universum geen saaie boekhouder is (Turing-machines zijn immers in eerste instantie bedoeld als modellen van saaie boekhouders). Een armzaliger wereldbeeld dan het laatste is nauwelijks denkbaar [Copeland and Sylvan, 1999].

Het staat echter zeker niet vast dat er super-Turing-machines bestaan of zelfs maar kunnen bestaan in de fysische realiteit. Het laatste woord over dit onderwerp is nog niet gezegd.

Ook al zouden er geen super-Turing-machines in de natuur voorkomen, dan zouden er misschien nog wel 'gewone' Turing-machines in de werkelijkheid kunnen bestaan. Er lijkt meer dan genoeg plaats te zijn voor zulke machines in een analoog universum. Een aantal jaren geleden heeft men bovendien laten zien dat Turing-machines op een heel natuurlijke manier in tamelijk 'gewone' dynamische systemen ingebed kunnen worden [Bennett, 1990, Bournez and Cosnard, 1996, Moore, 1990, 1991, 1998]. Misschien zijn we wel omringd door natuurlijke Turing-machines!

Wolfram [Wolfram, 1996] heeft deze gedachte aardig onder woorden gebracht:

> "We kunnen het gedrag van elk systeem in de natuur als een berekening opvatten: het systeem begint in een of andere toestand—dat is de input—gaat dan een tijdje zijn gang en belandt tenslotte in een of andere eindtoestand—die met de output correspondeert. Als, om een voorbeeld te geven, een vloeistof rond een obstakel stroomt, dan voert zij in feite een berekening uit over hoe ze moet stromen. Hoe gecompliceerd is die berekening? Het is zeker dat het ons veel moeite kost om het gedrag te reproduceren met de gebruikelijke wetenschappelijke rekenmethoden. Maar de grote ontdekking die ik heb gedaan is dat dit niet verrassend is: het natuurlijke systeem

> is in feite zelf een universele rekenmachine, en kan net zulke ingewikkelde berekeningen uitvoeren als welk ander systeem ook. Al onze mooie grote computers—net zoals trouwens onze hersenen—zijn met andere woorden dus niet in staat om berekeningen uit te voeren die ook maar iets gecompliceerder zijn dan de berekeningen die een hoeveelheid vloeistof kan uitvoeren. Het is een vernederende conclusie—als het ware de ultieme kleinering—na de ontdekking dat de aarde niet het centrum van het universum is, dat onze lichamen mechanisch werken, en zo voorts."

Als er Turing-machines in de natuur zouden bestaan, zou dat heel interessante gevolgen hebben. Zoals we hierboven bij het stop-probleem hebben gezien is het gedrag van Turing-machines op de lange termijn in principe niet voorspelbaar. Zal Turing-machine X gegeven input Y ooit stoppen? Er is doorgaans maar één manier om daar achter te komen: de machine simuleren. Maar er is geen garantie dat de simulatie ooit stopt! In overeenstemming hiermee, maar sterker nog, zegt de stelling van Rice dat iedere niet-triviale eigenschap van het gedrag van Turing-machines onbeslisbaar is [Hopcroft and Ullman, 1979, sec. 8.4]. Als er dynamische systemen in de natuur zijn die equivalent zijn met Turing-machines gelden deze resultaten natuurlijk ook voor deze systemen. Ook al ken je de begintoestand van zo'n systeem precies en is het systeem volstrekt deterministisch, er valt toch weinig over zijn toekomstige ontwikkeling te zeggen. Zal het systeem bijvoorbeeld ooit *chaotisch* worden in de zin van de chaos-theorie (dat wil zeggen: buitensporig gevoelig voor zelfs de miniemste veranderingen in de begintoestand)? Er zit in het algemeen weinig anders op dan maar af te wachten.

Gezien dit onvermogen om de toekomst te voorspellen van machines die equivalent zijn met Turing-machines, maakt het vanuit menselijk standpunt bekeken eigenlijk niet eens zo heel erg veel uit of er alleen maar Turing-machines of ook super-Turing-machines in de natuur bestaan: de wetenschap zal zelfs al in het eerste geval in voorspellende kracht tekort schieten.

Een belangrijke tegenoverweging tegen de mogelijkheid van het bestaan van Turing machines en super-Turing-machines in de fysische werkelijkheid berust op een theorema van Bekenstein, dat zegt dat ieder natuurkundig systeem slechts een eindige hoeveelheid informatie kan bevatten [Bekenstein, 1981]. 'Hoeveelheid informatie' betekent in dit geval 'aantal onderscheidbare quantumtoestanden'. Als gevolg van de onzekerheidsrelaties in de quantum-mechanica geldt dat de hoeveelheid informatie in een bolvormig gebied met massa M en straal R hooguit kMR bits bedraagt, waar k een constante is met een waarde van ongeveer 2.6 maal tien tot de drieënveertigste bits per kilogram maal meter. Deze 'Bekenstein-bovengrens' impliceert dat ieder deel van het heelal (dat immers slechts een eindige massa en eindige straal heeft) hooguit een eindige automaat is, geen Turing-machine of super-Turing-machine. (Turing machines kunnen in een oneindig aantal toestanden verkeren doordat ze een onbeperkte band hebben, super-Turing-machines kunnen dat *a fortiori* doordat hun 'ingebouwde orakel' een oneindige hoeveelheid informatie bevat.)

Een andere tegenwerping tegen de mogelijkheid van het bestaan van Turing machines en super-Turing-machines in de fysische natuur is het feit dat vele (maar niet alle) moderne natuurkundigen ervan overtuigd zijn dat de ruimtetijd in laatste instantie discreet, niet continu is [Gibbs, 1998, pp. 88–115]. Omdat

de ruimte-tijd begrensd is, zou dit impliceren dat ieder deel van het universum hooguit een eindige automaat vormt.

Een voorbeeld van een discrete aanpak in de fysica zijn de modellen waarin men de kosmos als een 'cellulaire automaat' beschouwt, dat wil zeggen, als een groot schaakbord met stukken die alleen in interactie staan met hun naaste buren en waarvan het gedrag wordt bepaald door eenvoudige regeltjes. De bekendste cellulaire automaat is *the game of life* van Conway [Poundstone, 1985]. Zelfs de simpelste regeltjes blijken al tot onverwacht grote emergente complexiteit te leiden. Zo kun je in *the game of life* een universele Turing-machine construeren, een Turing-machine die alle andere Turing-machines kan simuleren. Fysische fenomenen die op andere manieren nauwelijks te modelleren zijn, zoals turbulentie in vloeistoffen (denk aan het citaat van Wolfram hierboven), blijken op deze manier soms bijzonder aardig te kunnen worden gesimuleerd. Discrete fysici zijn erop uit om de hele fysica op deze manier te 'digitaliseren'. De bekendste namen op dit gebied zijn Fredkin, Landauer, Margolus, Toffoli en Wolfram. De filosoof Steinhart brengt hun ideeën tegenwoordig als 'digitale metafysica' aan de man [Steinhart, 1998].

Als de discrete fysici gelijk hebben, is ook de mens in laatste instantie natuurlijk een soort cellulaire automaat—een eindige, en daarom niet krachtiger dan een eindige automaat.

Wie hebben er gelijk, de aanhangers van de oude analoge school of de voorstanders van de nieuwe discrete school? Dit valt op dit moment niet uit te maken. De fysici zullen zelf het antwoord moeten geven.

Een moeilijkheid bij dit alles is nog wel dat we over ieder fysisch systeem altijd slechts een *eindig* aantal gegevens met een *eindige* precisie zullen kunnen bemachtigen. En aan zo'n beperkte voorraad gegevens kan altijd recht worden gedaan door een eindige automaat. Het is dus de vraag in hoeverre de hypothese dat systeem X een Turing-machine of super Turing-machine is empirische betekenis heeft [Nielsen, 1997, Ozawa, 1998]. (Iets dergelijks geldt ook voor de vraag of de ruimte-tijd discreet is [Forrest, 1995].) Het lijkt erop dat de knoop uiteindelijk zal moeten worden doorgehakt door theoretische overwegingen. De situatie is hierbij hetzelfde als toen men voor de keus stond de banen van de planeten als ellipsen of als epicykels op epicykels op ... op epicykels op cirkels te beschrijven. De observationele consequenties van beide benaderingen lopen niet ver uiteen [Hoyle, 1962, Appendix]. Maar de ellips-interpretatie was veel eleganter en hanteerbaarder en heeft daarom het pleit gewonnen. Zo zou het eventueel ook met de niet-eindige-automaat benadering kunnen gaan.

# 6 Taalkunde

Er is nog één wetenschap die van belang is bij het beantwoorden van de vraag in welke klasse van automaten de mens moet worden ondergebracht: de taalkunde.

Taalkundigen beweren vaak dat mensen geen eindige automaten zijn omdat sommige natuurlijke talen 'niet-contextvrij' zijn [Brandt Corstius, 1974]. Dit is echter geen goed argument. Een mens kan in zijn of haar korte leven immers hooguit een beheersing van een *eindig fragment* van een dergelijke niet-contextvrije taal ten toon spreiden. Maar dat kunnen eindige automaten ook [Lokhorst, 1991].

In verband met eindige automaten is het nog van belang om op te merken

dat sommige filosofen hebben beweerd dat de theorie dat de mens een eindige automaat is zinledig is. Putnam heeft beweerd dat "een steen iedere eindige automaat realiseert" [Putnam, 1988], Searle dat "de moleculen in mijn behang WordStar implementeren" [Searle, 1990]. We zullen hier niet op deze argumenten ingaan omdat anderen ze al heel goed weerlegd hebben [Chalmers, 1994a,b, 1996a,b, Copeland, 1996, O'Rourke and Shattuck, 1993].

# 7 Tussenbalans

We zijn allerlei soorten automaten tegengekomen. Sommige ervan zijn beduidend zwakker dan de Turing-machine (neurale netwerken met ruis, eindige automaten), andere precies even sterk, weer andere veel krachtiger dan de krachtigste Turing-machine. Al deze soorten automaten zouden, voor zover we thans weten, in principe in de natuur kunnen voorkomen. Waar moeten we de mens op deze schaal plaatsen? Niemand die het weet. Alle mogelijkheden liggen in principe nog open. Eenieder die van plan is de Turing-machine metafoor op de mens toe te passen zij dus gewaarschuwd. Op de wetenschap kan men zich hierbij in ieder geval niet beroepen.

Zelf neig ik om sentimentele redenen (bescheidenheid) naar het eindige automaten perspectief. Zoals gezegd is dit in principe toereikend. Waarom zouden we hiermee geen genoegen nemen?

# 8 Een praktischer perspectief

We kunnen ook een practischer perspectief kiezen en de zaak meer als ingenieurs bekijken. We zien dan dat de hersenen zo'n honderd miljard zenuwcellen bevatten. Elk daarvan is gemiddeld met zo'n duizend andere zenuwcellen verbonden. Per verbinding zullen er wel niet meer dan honderd bewerkingen per seconde worden uitgevoerd. Dit geeft ons een rekensnelheid van tien tot de zestiende bewerkingen per seconde, oftewel tien petaops. Ter vergelijking: er zijn tussen de tien tot de zestiende en tien tot de zeventiende mieren op aarde. De chips-fabrikant Intel maakte in 1997 ongeveer één transistor per mier [Moore, 1997].

Als de PC zich net zo snel blijft ontwikkelen als hij dat gedurende de afgelopen drie decennia heeft gedaan, zal er binnen een halve eeuw een betaalbare PC op de markt komen met een even grote rekensnelheid [Bell and Gray, 1997, Bostrom, 1998, Merkle, 1989, Moravec, 1998]. Als we aannemen dat de berekeningen die deze PC uitvoert vergelijkbaar zijn met de bewerkingen in de hersenen waarover we het hierboven hadden (we kiezen hiermee impliciet voor het eindige automaten perspectief op de hersenen), kan er dus over vijf decennia een betaalbare PC op de markt komen met een even grote rekenkracht als de hersenen.

Het menselijke geheugen lijkt al helemaal geen problemen op te leveren. Volgens allerlei schattingen slaat een mens gedurende zijn of haar hele leven maximaal een paar honderd megabytes op [Landauer, 1986, Merkle, 1988]. Dat past gemakkelijk op een CD-ROM of harde schijf van een computer.

Een dergelijke PC zal stellig nog geen menselijke intelligentie vertonen. Het lijkt al jarenlang zo te zijn dat de voortdurende verbetering van de hardware

teniet wordt gedaan door een voortdurende verslechtering van de software [Denning and Metcalfe, 1997*passim*]. (Vooral de gebruikers van Microsoft producten zullen deze stelling kunnen beamen.) Hoe beter de hardware, hoe moeilijker het kennelijk voor ons mensen valt om hem ten volle te benutten.

Er lijkt maar één manier te zijn waarop deze barrière kan worden overwonnen: door computers zelf te laten leren. Als het maken van kunstmatige intelligentie onze krachten te boven gaat, moet de computer het zelf maar doen. Turing zag dit al een halve eeuw geleden in [Turing, 1948, 1950]. Zijn eigen pogingen om tot lerende machines te komen waren nog niet erg succesvol [Copeland and Proudfoot, 1996, Turing, 1948]. Sindsdien is er echter veel vooruitgang geboekt. Men begint wegen te zien om het werk uit handen te geven. Te denken valt met name aan evolutionaire algoritmen (waarbij men zich laat inspireren door de evolutietheorie) en kunstmatige neurale netwerken. De spectaculairste successen komen uit de laatste hoek. Zo kan NETTalk, een eenvoudig éénrichtingsnetwerk met maar drie lagen zenuw-achtige cellen, iedere willekeurige taal in slechts één nacht correct leren uitspreken [Sejnowski and Rosenberg, 1987].

Er zijn dus perspectieven. We moeten met een tamelijk oningevulde, plastische structuur beginnen (voor ideeën hierover, zie [Franklin, 1995]) en de computer verder zelf het werk laten doen. Sommige onderzoekers zijn nu reeds aan het nadenken over het ontwikkelen van lesmateriaal voor de zelf-lerende computers van de nabije toekomst [Valiant, 1998]. Turing heeft het zelf trouwens ook al gehad over het uitdelen van straffen en beloningen aan computers om hun gedrag te corrigeren [Turing, 1948].

Als deze ontwikkelingen doorgaan, wat voor computer zal er dan uiteindelijk op mijn bureau staan (aannemende dat ik nog een aantal eeuwen te leven heb)? Daar valt op het moment nog niet veel van te zeggen. Of toch. De mogelijkheden zijn al aardig in kaart gebracht in de science fiction van de afgelopen eeuw. Zal mijn PC zijn omgeving met zijn electronische retina over een millenium even argwanend in de gaten houden als HAL dat deed in *2001*?

De interessantste vraag is misschien deze: zullen dergelijke toekomstige electronische breinen werkelijk kunnen zien, denken en voelen? Zullen ze een kleurrijk innerlijk leven hebben en zich daarvan bewust zijn? Er lijkt geen enkele reden te verzinnen te zijn waarom dat niet zo zou kunnen zijn [Lycan, 1998]. Er zijn wel 'protoplasma-chauvinisten' die er een andere mening op na houden, maar zij hebben bij mijn weten nog nooit een overtuigende reden voor hun standpunt gepresenteerd.

Hoe dit ook zij, het lijdt geen twijfel dat de computer ons nog lang te denken zal geven. Net zoals de klok en de stoommachine dat vroeger waren is de computer tegenwoordig een geweldige inspiratiebron voor filosofen en het ziet er naar uit dat hij dat nog heel lang zal blijven [Bynum and Moor, 1998].

Met dank aan George Berger, David Chalmers, Maarten Coolen, Jack Copeland, Jim Moor, Wolfgang Maass, Ellie Rhebergen en Ronald de Wolf voor hun nuttige opmerkingen en hun rol als klankbord.

# Referenties

J. D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Physical Review*, D23: 287–298, 1981.

G. Bell and J. N. Gray. The revolution yet to happen. In Denning and Metcalfe [1997].

C. H. Bennett. Undecidable dynamics. *Nature*, 346: 606–607, 1990.

L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag, Berlin, 1997.

Nick Bostrom. How long before superintelligence? http://www.hedweb.com/nickb/superintelligence.htm, 1998.

O. Bournez and M. Cosnard. On the computational power of dynamical systems and hybrid systems. *Theoretical Computer Science*, 168: 417–459, 1996.

H. Brandt Corstius. *Algebraïsche taalkunde*. Oosthoek, Utrecht, 1974.

A. W. Burks. Logic, computers and men. *Proceedings and Addresses of the American Philosophical Association*, 46: 39–57, 1973.

Terrell Ward Bynum and James H. Moor, editors. *The Digital Phoenix: How Computers Are Changing Philosophy*. Blackwell Publishers, Oxford (UK), 1998.

D. J. Chalmers. A computational foundation for the study of cognition. Philosophy-Neuroscience-Psychology Technical Report 94-03, Washington University, 1994a.

D. J. Chalmers. On implementing a computation. *Minds and Machines*, 4: 391–402, 1994b.

D. J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Oxford, 1996a.

D. J. Chalmers. Does a rock implement every finite-state automaton? *Synthese*, 108: 309–333, 1996b.

B. J. Copeland. What is computation? *Synthese*, 108: 335–359, 1996.

B. J. Copeland. The broad conception of computability. *American Behavioral Scientist*, 40: 690–716, 1997a.

B. J. Copeland. The Church-Turing thesis. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/, 1997b.

B. J. Copeland and D. Proudfoot. On Alan Turing's anticipation of connectionism. *Synthese*, 108: 361–377, 1996.

B. J. Copeland and R. Sylvan. Beyond the Universal Turing Machine. *Australasian Journal of Philosophy*, 77: 46–66, 1999.

N. C. A. Da Costa and F. A. Doria. Undecidability and incompleteness in classical mechanics. *International Journal of Theoretical Physics*, 30: 1041–1073, 1991.

D. C. Dennett. *Consciousness Explained*. Little, Brown, Boston, 1991.

P. J. Denning and R. M. Metcalfe, editors. *Beyond Calculation: The Next Fifty Years of Computing*. Springer-Verlag, Berlin, 1997.

P. Forrest. Is space-time discrete or continuous?–an empirical question. *Synthese*, 103: 327–354, 1995.

S. Franklin. *Artificial Minds*. MIT Press, Cambridge (Mass.), 1995.

R. Gandy. Church's thesis and principles for mechanisms. In J. Barwise, H. J. Keisler, and K. Kunen, editors, *The Kleene Symposium*. North-Holland, Amsterdam, 1980.

F. H. George. Communication to the editor. *Philosophy*, 32: 168–169, 1957.

Philip Gibbs. *Event-Symmetric Space-Time*. Weburbia Press, Bristol, 1998.

J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading (Mass.), 1979.

Fred Hoyle. *Astronomy*. Macdonald, London, 1962.

William James. *The Principles of Psychology*. Henry Holt and Co., New York, 1890.

S. C. Kleene. Representation of events in nerve nets and finite automata. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*, pages 3–41. Princeton University Press, Princeton (N. J.), 1956.

G. Kreisel. A notion of mechanistic theory. *Synthese*, 29: 11–26, 1974.

G. Kreisel. Review of Pour-El and Richards [1979] and Pour-El and Richards [1981]. *Journal of Symbolic Logic*, 47: 900–902, 1982.

T. K. Landauer. How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, 10: 477–493, 1986.

W. J. Levelt. Are multilayer feedforward networks effectively Turing machines? *Psychological Research*, 52: 153–157, 1990.

G. J. C. Lokhorst. Is de mens een eindige automaat? In F. Geraedts and L. de Jong, editors, *Ergo cogito III*. Historische Uitgeverij, Groningen, 1991.

J. R. Lucas. *The Freedom of the Will*. Clarendon Press, Oxford, 1970.

A. Lugg. Finite automata and human beings. In M. H. Salmon, editor, *The Philosophy of Logical Mechanism*. Reidel, Dordrecht, 1990.

William G. Lycan. Qualitative experience in machines. In Bynum and Moor [1998].

W. Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10: 1659–1671, 1997.

W. Maass and C. M. Bishop, editors. *Pulsed Neural Networks*. MIT Press, Cambridge (Mass.), 1999.

W. Maass and P. Orponen. On the effect of analog noise in discrete-time analog computations. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge (Mass.), 1997.

W. Maass and E. D. Sontag. Analog neural nets with Gaussian or other common noise distributions cannot recognize arbitrary regular languages. *Neural Computation*, 11: 771–782, 1999.

W. S. McCulloch. *Embodiments of Mind*. MIT Press, Cambridge (Mass.), 1965.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5: 115–133, 1943. Reprinted in McCulloch [1965].

B. Meltzer and D. Michie, editors. *Machine Intelligence*, volume 5. Edinburgh University Press, Edinburgh, 1969.

R. C. Merkle. How many bytes in human memory? *Foresight Update*, (No. 4), October 1988.

R. C. Merkle. Energy limits to the computational power of the human brain. *Foresight Update*, (No. 6), August 1989.

J. H. Moor. Thinking must be computation of the right kind. In *Proceedings of the 20th World Congress of Philosophy*, 1998a. To appear.

James H. Moor. Assessing artificial intelligence: Chess and the Turing test. In Bynum and Moor [1998].

C. Moore. Unpredictability and undecidability in dynamical systems. *Physical Review Letters*, 64: 2354–2357, 1990.

C. Moore. Generalized shifts: Unpredictability and undecidability in dynamical systems. *Nonlinearity*, 4: 199–230, 1991.

C. Moore. Finite-dimensional analog computers: Flows, maps, and recurrent neural networks. In C. S. Calude, J. Casti, and M. Dinneen, editors, *Unconventional Models of Computation*. Springer-Verlag, Singapore, 1998.

G. Moore. An update on Moore's law. INTEL Developer Forum Keynote, San Francisco, September 30, 1997.

H. Moravec. When will computer hardware match the human brain? *Journal of Transhumanism*, 1, 1998.

R. J. Nelson. *The Logic of Mind*. Reidel, Dordrecht, 2nd edition, 1989.

M. A. Nielsen. Computable functions, quantum measurements, and quantum dynamics. *Physical Review Letters*, 79: 2915–2918, 1997.

J. O'Rourke and J. Shattuck. Does a rock realize every finite automaton? A critique of Putnam's theorem. Technical Report 30, Department of Computer Science, Smith College, Northampton (Mass.), 1993.

Masanao Ozawa. Measurability and computability. http://xxx.lanl.gov/abs/quant-ph/9809048, 1998.

R. Penrose. *The Emperor's New Mind*. Oxford University Press, Oxford, 1989.

R. Penrose. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press, Oxford, 1994.

W. Poundstone. *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. Morrow, New York, 1985.

M. B. Pour-El. Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers). *Transactions of the American Mathematical Society*, 199: 1–28, 1974.

M. B. Pour-El and I. Richards. A computable ordinary differential equation which possesses no computable solution. *Annals of Mathematical Logic*, 17: 61–90, 1979.

M. B. Pour-El and I. Richards. The wave equation with computable initial data such that its unique solution is not computable. *Advances in Mathematics*, 39: 215–239, 1981.

M. B. Pour-El and I. Richards. Noncomputability in models of physical phenomena. *International Journal of Theoretical Physics*, 21: 553–555, 1982.

H. Putnam. Minds and machines. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. Brains and behavior. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. Robots: Machines or artificially created life? In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. The mental life of some machines. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. The nature of mental states. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. Logical positivism and the philosophy of mind. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. Philosophy and our mental life. In *Mind, Language, and Reality* Putnam [1975].

H. Putnam. *Mind, Language, and Reality*. Cambridge University Press, Cambridge, 1975.

H. Putnam. *Representation and Reality*. MIT Press, Cambridge (Mass.), 1988.

H. Putnam. The project of artificial intelligence. In *Renewing Philosophy*, pages 1–18. Harvard University Press, Cambridge (Mass.), 1992.

H. Putnam. Persoonlijke mededeling, Rotterdam, november 1995. 1995.

G. Rey. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Basil Blackwell, Oxford, 1997.

L. A. Rubel. The brain as an analog computer. *Journal of Theoretical Neurobiology*, 4: 73–81, 1985.

L. A. Rubel. Digital simulation of analog computation and Church's thesis. *Journal of Symbolic Logic*, 54: 1011–1017, 1989.

G. H. E. Russelmann. *Van James Watt tot Sigmund Freud*. Van Loghum Slaterus, Deventer, 1983.

B. Scarpellini. Zwei unentscheitbare Probleme der Analysis. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9: 265–289, 1963.

J. R. Searle. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64: 21–37, 1990.

T. Sejnowski and C. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1: 145–168, 1987.

H. T. Siegelmann. *Foundations of Recurrent Neural Networks*. PhD thesis, Rutgers University, New Brunswick (N. J.), 1993.

H. T. Siegelmann. Computation beyond the Turing limit. *Science*, 268: 545–548, 1995.

H. T. Siegelmann. *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhauser, Boston, 1998.

H. T. Siegelmann and E. D. Sontag. Analog computation via neural networks. *Theoretical Computer Science*, 131: 331–360, 1994.

H. T. Siegelmann and E. D. Sontag. Computational power of neural networks. *Journal of Computer System Sciences*, 50: 132–150, 1995.

M. Stannett. X-machines and the halting problem: Building a super-Turing machine. *Formal Aspects of Computing*, 2: 331–341, 1990.

Eric Steinhart. Digital metaphysics. In Bynum and Moor [1998].

I. Stewart. Deciding the undecidable. *Nature*, 352: 664–665, 1991.

Karl Svozil. The Church-Turing thesis as a guiding principle for physics. http://xxx.lanl.gov/abs/quant-ph/9710052, 1997.

A. M. Turing. On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society, Series 2*, 42: 230–265, 1936.

A. M. Turing. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, 45: 161–228, 1939.

A. M. Turing. Intelligent machinery. Technical report, National Physical Laboratory, 1948. Reprinted in Meltzer and Michie [1969].

A. M. Turing. Computing machinery and intelligence. *Mind*, 59: 433–460, 1950.

L. G. Valiant. A neuroidal architecture for cognitive computation. Technical Report TR-04-98, Computer Science Group, Harvard University, 1998.

Tim van Gelder. What might cognition be, if not computation? *The Journal of Philosophy*, 91: 345–381, 1995.

Tim van Gelder. The dynamical hypothesis in cognitive science. *The Behavioral and Brain Sciences*, 21: 615–628, 1998.

A. Vergis, K. Steiglitz, and B. Dickinson. The complexity of analog computation. *Mathematics and Computers in Simulation*, 28: 91–113, 1986.

P. Vroon and D. Draaisma. *De mens als metafoor*. Ambo, Baarn, 1985.

S. Wolfram. Undecidability and intractability in theoretical physics. *Physical Review Letters*, 54: 735–738, 1985.

S. Wolfram. Computers, science, and extraterrestrials: An interview with Stephen Wolfram. In D. G. Stork, editor, *HAL's Legacy: 2001's Computer as Dream and Reality*. MIT Press, Cambridge (Mass.), 1996.