Bas Haring – De IJzeren Wil

Joris de Ruiter,
tweedejaars student AI aan de Vrije Universiteit

‘Veel mensen hebben het gevoel dat wij iets extra’s hebben ten opzichte van dieren en machines: een ziel, bewustzijn, of een wil. Maar is dat gevoel terecht?’. Deze vraag probeert Bas Haring te beantwoorden in zijn boek ‘De IJzeren Wil’. In deze eerste LiterAlteiten wordt Bas Harings boek door een kritische AI-bril bekeken: interessant of gesneden koek?

Hersenen vs. computers
De hersenen bestaan uit een netwerk van ongeveer 10^{10} domme neuronen met daartussen in totaal 10^{14} verbindingen. Het brein is geen goed georganiseerde fabriek, het heeft meer iets weg van een mierenhoop. Het bestaat uit vele domme, onwetende wezens (neuronen) die niet altijd even goed samenwerken. Het nadeel hiervan is inefficiëntie. Het voordeel is flexibiliteit en plasticiteit.

Computers hoeven niet elektrisch te zijn. Een draaiorgel (Turing machine à la Bas Haring) voldoet ook. Computers bestaan uit een geheugen en een processor. Elk programma is op elke computer te draaien, mits er een werkende versie voor die specifieke computer wordt geschreven en je langzame computers lang genoeg de tijd geeft.

In tegenstelling tot hersenen zijn computers strak gestructureerd, snel en efficiënt. ‘Een computer is qua organisatie haast het tegenovergestelde van een brein: één razendsnelle, strak georganiseerde rekeneenheid, tegenover een paar miljard trage, door elkaar heen kakelende hersencellen.’

Kan een machine slim zijn?
Een kenmerk van intelligentie is logisch redeneren. Wij mensen kunnen dat redelijk, andere dieren kunnen het niet, en computers kunnen het beter. Mensen zouden nu kunnen zeggen: ‘Maar een computer blijft stom: hij draait enkel instructies af’. In een reactie hierop komt Bas Haring met een superintelligente wetenschapper van Mars. Deze doorziet en begrijpt de hersenen zo goed, dat deze kan concluderen dat de hersenen ook maar een programma afdraaien. Als we de computer stom vinden om het feit dat hij ‘slechts’ een programma afdraait, zouden we onszelf ook stom moeten vinden. En aangezien we dat niet graag doen, lijkt het hem onnodig hier naar te kijken:

‘Het feit dat je precies weet wat er in een computer gebeurt, doet niets af aan zijn intelligentie. Mocht er een Marsman bestaan die precies weet wat er in jouw brein afspeelt, dan doet dat immers ook niks af aan jouw intelligentie.’

Liever kijkt hij naar wat we eigenlijk kunnen. En als we kijken naar redeneren, moeten we toegeven dat computers daar beter in zijn dan wij. Blijft het feit dat redeneren niet de gehele lading dekt van ‘intelligentie’ of ‘slim zijn’.

Woorden, wil, bewustzijn
Na een aantal van deze vragen te hebben beantwoord, concludeert hij dat computers aardig wat vaardigheden hebben. Een reactie van vele mensen zou kunnen zijn: ‘Maar een computer blijft stom: het draait enkel instructies af. Wij hebben iets extra’s. Toch?’ Bas Haring gaat deze extra’s vervolgens langs: leven, voelen, jezelf kennen, willen, bewustzijn, denken, begrijpen, intelligent zijn, etc. Voor elk extraatje dat wij mensen onszelf toekennen, heeft hij wel een metafoor klaar om het te ontkrachten. Hieronder zullen er enkele volgen.

Allereerst zullen we echter kijken naar de basis: woorden. Volgens Haring kun je veel over die extraatjes zeggen, maar in ieder geval geldt dat het woorden zijn, door mensen gemaakt en gebruikt. Daarbij geven we woorden aan materiële en immateriële zaken. Materieel als in ‘de hersenen’, immaterieel als in ‘denken, voelen, willen, etc.’. Volgens Haring zijn beide even echt, maar toch kennen vele mensen die immateriële zaken alleen toe aan mensen, in mindere mate aan dieren, en zeker niet aan apparaten. Hoe komt dat?

Volgens Haring komt dat omdat we de woorden gebruiken om de wereld en onszelf te begrijpen, maar dan alleen als we geen logische verklaring kunnen vinden. Zo kunnen we bijvoorbeeld kijken naar een geprogrammeerd computervisje. Het visje wordt in een virtueel aquarium gezet en moet vervolgens overleven tussen een stel hongerige haaien.

Als we in het visje een aantal simpele regels programmeren, zal het makkelijk zijn het gedrag van de vis terug te voeren op die regels. Zien we bijvoorbeeld het visje naar links gaan, dan kunnen we in de code vinden dat hij dat doet als een haai rechts van hem te dichtbij komt. Er lijkt weinig aanleiding te zijn het visje een vrije wil toe te kennen: hij gaat gewoon naar links omdat dat in zijn code staat. We kunnen die code ook prima aanwijzen: if(shark = right){swim to left}

Laten we echter zijn regels bepalen door de evolutie (1000 visjes, de beste overleven en paren, nieuwe visjes erven telkens een helft van de code van beide ouders, etc) dan komen we uit

op een werkelijk ongestructureerd, onlogisch en onhandig stuk code, ware het niet dat het werkt (het visje overleeft). Stel dat we nu het visje naar links zien gaan, hoe kunnen we dat dan verklaren? Je kunt het hem niet vragen, en het ook niet in zijn broncode vinden. Er blijft weinig anders over dan te zeggen dat het visje blijkbaar naar links 'wilde'.

Zie daar de kern van 'de wil': we hebben hem alleen nodig als het te complex wordt. En volgens Bas Haring geldt dat voor al die 'extraatjes': het zijn woorden die we gebruiken om onszelf te begrijpen. Daarbij zijn wijzelf zo complex, zijn dieren het in mindere mate, en zijn machines het vaak nog niet. Maar voor hoe lang?

Over bewustzijn zegt Bas Haring ongeveer het volgende. Wij mensen vinden dat we bewustzijn hebben omdat we dat zo voelen. We zijn echter goedgelovig: we voelen en geloven wel vaker dingen waarvan we nog niet hebben bewezen dat ze er daadwerkelijk zijn (bijvoorbeeld religieuze zaken). Hij laat daarom menselijke gevoelens en emoties voor wat ze zijn (een interne beleving), en kijkt puur naar de buitenkant. En van de buitenkant gezien kan hij niets anders concluderen dan dat bewustzijn een woord is dat wij mensen gebruiken om ons gedrag te benoemen. Stel dat er een robot komt die zich net zo gedraagt als wij, wat hebben wij dan voor reden onszelf wel bewustzijn toe te kennen en de robot niet?

Stijl

Het boek is geschreven voor een breed publiek, en dat is te merken. Hij heeft het niet over 'neurale netwerken', maar over 'een grote touwtrekkerij van hersencellen', en over allerlei 'landen' (hersenkwabben) die andere 'landen' inlichten als er gevaar dreigt. 'Exciteren' en 'inhiberen' worden 'schouderklopjes' en 'reprimandes'. Er wordt veel gebruik gemaakt van metaforen en alledaagse voorbeelden.

Of dit voor de gevorderde AI'er bevalt, ligt er maar aan. Je zou je kunnen ergeren aan de vele versimpelende voorbeelden en de vele woorden die hij nodig heeft om iets uit te leggen. Aan de andere kant kun je je ook verbazen over de goed gevonden metaforen en deze misschien gebruiken om niet- AI'ers uit te leggen hoe de hersenen werken.

Inhoudelijk gezien worden er al met al een aantal interessante dingen behandeld, bijvoorbeeld over de wil, emotie, zelf organiserende systemen, leren, denken en de betekenis van woorden.

Ikzelf vond het een interessant boek met een aantal goede ideeën, maar ik zou het persoonlijk niet kopen. Mocht je het een keer tegenkomen bij een vriend of in de bieb: lees het gerust. Je hebt het in een paar dagen uit. ϕ

Titel: De ijzeren wil
Auteur: Bas Haring
ISBN: 9052407207
Uitgever: Houtekiet
Prijs: □ 15,95

WWW.CONNECTIE.ORG